RADC-TR-80-119
Final Technical Report
May 1980

ADVANCED PARALLEL PROCESSING TECHNIQUES FOR I&W

Measurement Concept Corporation

Fred Tims
Penny Smith
Garry Caron
Hank Spaaenburg

LEVEL

ROME AIR DEVELOPMENT CENTER
Air Force Systems Command
Griffiss Air Force Base, New York 13441

AD A088421

DTIC
SELECTE
AUG 2 6 1980
C

80 8 25 001

This report has been reviewed by the RADC Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

Appendix A consists of Sections 3.0 through 3.5.

RADC-TR-80-119 has been reviewed and is approved for publication.

APPROVED: *John E. Frank*

JOHN E. FRANK
Project Engineer

APPROVED: *Thadeus J. Domurat*

THADEUS J. DOMURAT, Acting Chief
Intelligence & Reconnaissance Division

FOR THE COMMANDER: *John P. Huss*

JOHN P. HUSS
Acting Chief, Plans Office

(19) REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER RADC-TR-80-119	2 GOVT ACCESSION NO. AD-A088 451	3 RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) ADVANCED PARALLEL PROCESSING TECHNIQUES FOR I&W		5 TYPE OF REPORT & PERIOD COVERED Final Technical Report
		6 PERFORMING ORG REPORT NUMBER N/A
7. AUTHOR(s) Fred Tims Garry Caron Penny Smith Hank Spaaenburg		8 CONTRACT OR GRANT NUMBER(s) F30602-78-C-0190
9. PERFORMING ORGANIZATION NAME AND ADDRESS Measurement Concept Corporation 1721 Black River Blvd Rome NY 13440		10 PROGRAM ELEMENT PROJECT TASK AREA & WORK UNIT NUMBERS 19550403 64750F
11. CONTROLLING OFFICE NAME AND ADDRESS Rome Air Development Center (IRDA) Griffiss AFB NY 13441		12 REPORT DATE May 1980
		13 NUMBER OF PAGES 173
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Same		15 SECURITY CLASS (of this report) UNCLASSIFIED
		15a DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Same

18. SUPPLEMENTARY NOTES

RADC Project Engineer: John E. Frank (IRDA)

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

PDSC	DBMS-11
ADCOM	SARP
Parallel Processing	
Modeling (Analytic & Simulation)	

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report summarizes work done to predict the performance of two target computer systems (ADCOM & PDSC) under varying loads and configurations. Interactive FORTRAN programs have been developed to predict performance of various hardware configurations for crisis workloads. New technology looked at for inclusion in various hardware configurations is included as Appendix A.

DD FORM 1 JAN 73 1473   EDITION OF 1 NOV 65 IS OBSOLETE

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

LIST OF FIGURES

EVALUATION

This report summarizes work done to predict the performance of two
target computer systems (ADCOM & PDSC) under varying loads and configurations.
Interactive FORTRAN programs have been developed to predict performance of
various hardware configurations for crisis workloads. New technology looked
at for inclusion in various hardware configurations is included as
Appendix A.

JOHN E. FRANK
Project Engineer

1.0 INTRODUCTION

1.1 Purpose

This report describes the work performed by Measurement Concept Corporation (Mc^2), under Contract No. F30602-78-C-0190, to establish a methodology for evaluating available and projected parallel processing technology, as applicable, to cost effectively increase the capabilities of current and future intelligence systems. It is assumed that these intelligence systems are composed of a network of computers with associated software, terminals, and personnel to support intelligence analysis and production functions. The intelligence network may be organized into a number of subsystems to provide message handling, network control, user support, data base management, program development, and graphics terminal support functions, among others.

1.2 Executive Summary

1.2.1 Interpretation of Precedence Chart

Figure 1-1 shows graphically the precedence relationships of the technological developments which contribute to improved intelligence services. The connections between boxes are interpreted as follows:

o   All vertical lines meeting a horizontal line from the top are assumed to be connected to all vertical lines meeting the horizontal line from the bottom.

o   Reading up from any box, the heavy lines indicate developments which are either required as precedents for the development in the chosen box, or without which the development is only marginally attractive. The
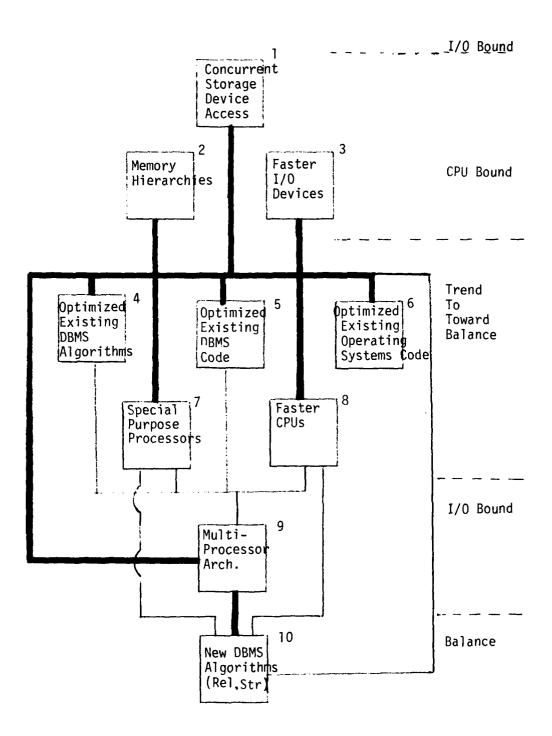
Figure 1-1   Advanced Technology Precedence Chart

paths with lighter lines connect developments which can
contribute to a system which includes the development
under question. For example, the bottom box, "New DBMS
Algorithms", requires its immediate predecessor, "Multi-
Processor Architectures", and its attractiveness is
increased by the development of faster raw data access
represented in the top three boxes. "Faster CPUs" and
"Special Purpose Processors" can enhance such a system.

1.2.2 Discussion - Overview

Historically, data base operations have been I/O bound, that is, they
have been limited in their data delivery and update rate by secondary
storage devices which have been significantly slower than the central
processing hardware and software which has driven them. Concurrent
storage device access, made possible by re-entrant, multi-threaded
software, along with memory hierarchy technology and emerging faster
storage devices will soon bring down I/O service time to the point
where it is less than the CPU time being used to control and process
it. At this point the system becomes CPU bound.

The advent of higher retrieval rates from secondary storage has made
the efficiency of DBMS software more critical. Previously, DBMSs
have been implemented with the primary efficiency consideration being
that of minimizing the number of I/Os which are performed as the
result of a request. The efficiency of the DBMS code and internal
algorithms and procedures has been considered in many, if not most,
cases relatively unimportant due to its being masked by comparably
high I/O service times. In some cases, moreover, DBMSs have been
implemented which are CPU bound even now, before the realization of
improved I/O rates. These DBMSs were developed with the primary
goal being that of providing powerful, highly generalized services,

with little care given to minimizing utilization of the CPU. For
these reasons, we urge that serious consideration be given to examining
off-the-shelf DBMSs being installed in intelligence systems, with a
view to improving both the internal procedures and the way in which
they have been coded. This would be a near-term, low cost effort
which would possibly provide significant improvement in systems product-
ivity and response. Any improvements in this area could carry forward
to the intermediate term and can apply to some forms, but not all,
of multi-processor configurations. They do not apply to long-term
solutions involving advanced DBMS development (see Figure 1-1).

Another near term enhancement effort, related to the above, is the
optimization of existing operating systems. That there is room for
improvement in operating systems for the 21(V) line, RSX-11D and,
probably, IAS, is evident in that RSX-11M incorporated just such
improvements. Short of moving to RSX-11M, such improvements are
considered non-cost-effective at this stage of systems development
for the following reasons:

    o   The technical risk of modifying these operating systems,
         with subsequent heavy validation requirements is high.

    o   These operating systems will not apply to intermediate and
         longer term system multi-processor configurations.

Care should be taken, however, that efficiency of algorithms and
implementation code be high on the list of criteria for new
operating systems.

Faster Central Processing Units exist today than are scheduled for
initial installation in the two systems studied on this project. In
some cases we have recommended their substitution. The long-term
potential for improvement in CPU speeds is limited by laws of physics.

Further improvements in processing speeds will come from the
introduction of Special Purpose Processors employed as functionally
dedicated auxilliaries of the general purpose processor(s) and from
the development of multi-processor architectures. Special purpose
processors are designed to perform specific functions such as array
comparisons (associative memory processors) or array manipulations
(vector machines).

Multi-processor architectures provide the long-term solution to future
requirements as well as more near-to-intermediate-term improvement.
This concept covers a wide range of realized and realizable implement-
ations, from those that are near term and limited, such as the
functionally distributed configurations in both of the subject systems,
to more efficient but still limited "pooled CPU" configurations wherein
a limited number of CPUs share both load and memory, to more advanced
arrays of general and special purpose processors combined with in-
expensive and highly efficient micro-processors. We have recommended for
some nodes of the subject systems the initial installation of limited,
pooled CPU machines. We have also recommended that effort continue
on the development of the more powerful multi-processor systems.

With the introduction of the more advanced multi-processor configura-
tions and the anticipated explosion of requirements, discussed in this
report, the intelligence systems of the future will be, once again, I/O
bound unless steps are taken to provide an adequate data delivery rate.

We have recommended for future, next generation intelligence data
handling systems, that steps be taken to implement relationally based
DBMSs supported by the data "streaming" approach to data access. The
relational model of data provides not only enhanced logical capabilities
for data retrieval and manipulation, but also reduced DBMS complexity

and overhead. Data streaming, wherein data is read and processed physically sequentially, avoids mechanical delays. This approach is recommended due to our doubts as to the cost-effective availability of large non-mechanical memories by the time-frame required. In addition, as larger portions of the data base are accessed as the result of requests for massive correlation functions or extensive, comprehensive displays involving, in some cases, the entire data base, the streaming technique will become more attractive than "search" techniques, even assuming the availability of multi-billion character non-mechanical memories.

Figure 1-2 shows the same relationships shown in Figure 1-1 projected on a time-line. The dates indicated are approximate and indicate the anticipated time of installation of the technologies in operational intelligence data handling systems.
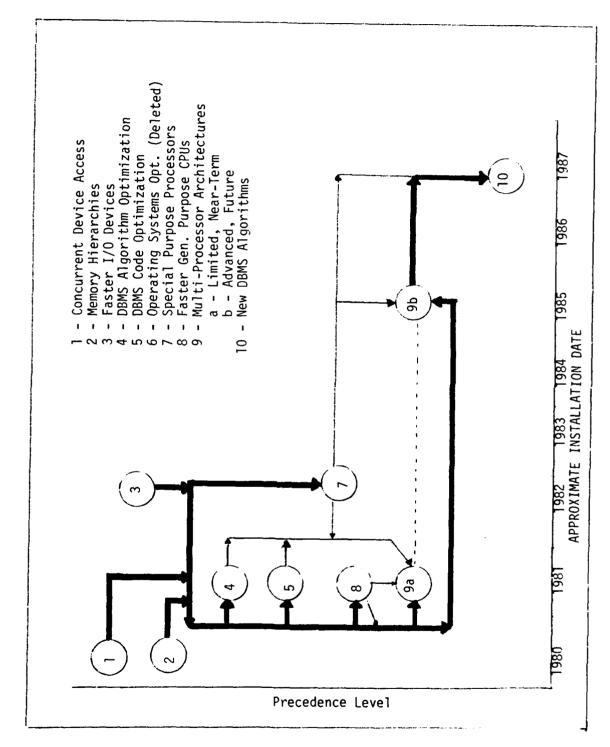
Figure 1-2   Enhancement Time-Line

1 - Concurrent Device Access
2 - Memory Hierarchies
3 - Faster I/O Devices
4 - DBMS Algorithm Optimization
5 - DBMS Code Optimization
6 - Operating Systems Opt. (Deleted)
7 - Special Purpose Processors
8 - Faster Gen. Purpose CPUs
9 - Multi-Processor Architectures
     a - Limited, Near-Term
     b - Advanced, Future
10 - New DBMS Algorithms

Precedence Level

1980   1981   1982   1983   1984   1985   1986   1987
APPROXIMATE INSTALLATION DATE

1.2.3 Discussion - Detail

The conclusions of this investigation fall into two categories,
each requiring a different treatment.

1.2.3.1 The Next Generation

The first is a consideration of the processing power, data structures,
and system architectures which will be required to satisfy require-
ments of a heretofore impossible level. These requirements are
those deriving from the implementation of dramatically more power-
ful and useful intelligence products. Among these are included
mobile target tracking, real-time sensor input and update, advanced
correlation and inference techniques, and more complex and resource
consuming graphic aids.

These functions are, as yet, unquantified as to their demands on a
computer system. It can only be surmised that they will require two or
more orders of magnitude greater processing power. For instance,
event correlation based on ad hoc matching of essential elements of
information with history or entity descriptor files will require
that all occurrences of selected data elements of all pertinent files,
containing billions of characters, be examined in a few seconds. Adding
to this single example an environment in which perhaps hundreds of
analysts are using the system for similar operations, it can be
perceived that current approaches can not provide even the gross
throughput to satisfy the requirement.

Such services are considered "requirements" in this study for the
simple reason that they will soon be possible. They are here dealt
with as "future" systems as opposed to "near future". They are seen
as evolving in parallel with the current generation of intelligence
systems. "Near future" here means current types of services provided
in much larger volumes, for more users, thus requiring enhancements of
current implementations. The future systems will be providing radically
advanced services and will require a revolution in technology. That
revolution has begun and is typified by current work in new processor
architectures and data base machines. This new technology is examined
in Volume I, Section 3 of the System/Subsystem Specification
produced under this contract. To attempt to apply this just emerging
technology to as yet unquantified requirements was considered to be
an overly speculative endeavor, considering the scope and pragmatic
intent of this project.

1.2.3.2 The Current Generation

Effort during this project has been concentrated in supplying enhance-
ment recommendations and development scenarios for current generation
intelligence systems (i.e., ADCOM and PDSC) under ever increasing

traffic loads. Some of the answers for the long term for these
systems will begin to utilize the advanced technology required for
the next generation. This technology includes multi-processor
control schemes, special purpose processors, faster peripheral
storage, and data base machines.

The following sections of this report summarize the accomplishments
of this project in developing a methodology for applying advanced
technology to intelligence systems at all levels, and in exercizing
this methodology against the current and projected ADCOM and PDSC
systems. A more detailed discussion of the methodology (character-
ization, modelling, technology application) can be found in Volume I
of [MEAS79]. Detailed results of the application of the methodology,
as well as specific recommendations, can be found in Volumes II
(ADCOM) and III (PDSC) of the same report. Somewhat more general
treatments are included in this report.

1.2.3.3 Investigation Findings

An overall summary of the findings of this study may be built about
a framework of successive bottleneck discovery and elimination. Minor
bottlenecks such as requirements for more analyst or quality control
terminals, or faster communications lines will not be summarized. Their
elimination is rather straightforward. This summary will deal with
the more serious bottlenecks created by the actual obtaining of data
and processing of that data.

1.2.3.3.1 Bottleneck #1 HOST Central Processor

It was found that, in ADCOM, the CPU specified for the HOST node
(HIS 6068), rated at .55 MIPS (million instructions per second),
would be inadequate to perform both data base operations and

applications processing. The preferred solution was to break the
HOST node into two nodes, as is being implemented for the PDSC
system - one processor for data base services and one for applications
processing. This separation of functions, in addition to providing
adequate CPU resources, lends a desirable flexibility to the system,
making future enhancement efforts more feasible in that they can be
focused more accurately on the problem at hand.

1.2.3.3.2 <u>Bottleneck #2 Data Base Management System</u>

1.2.3.3.2.1 <u>Single Threading</u>

This bottleneck is in software rather than hardware. It consists of
a DBMS which is single rather than multithreaded. Simply stated, a
single-thread DBMS will process requests one at a time, performing
all I/Os for a request, meanwhile suspending its CPU processing
during I/O activity, before it will accept another request. Activity
is sequential - DBMS related CPU processing occurs while the disks
are idle, and disk seeks occur while the DBMS related CPU processing
is suspended. If CPU processing associated with an I/O is 13 ms
and disk time is 40 ms, a maximum of approximately 68.000 I/Os can
be performed in an hour. This just barely meets initial ADCOM
throughput requirements of 65,000 to 70,000 I/Os per hour and falls
far short of PDSC initial requirements of approximately 300,000 I/Os
per hour. Response times even for the ADCOM load will be unacceptable.

A multi-threaded DBMS can process several requests concurrently,
overlapping disk seek times with each other, if they occur on different
disks, and overlapping CPU time with disk time. A multi-threaded
DBMS can theoretically support a maximum of approximately 260,000 I/Os
per hour given the service times mentioned above, and assuming

that an average of three disks are being driven concurrently for an average effective access time of 14 ms (40 ms ÷ 3 = 13 1/3 ms, overlapped 100% with CPU overhead of 13 ms, plus 2/3 ms transfer time). For higher levels of disk concurrency (determined by physical data organization and user request patterns), or for more efficient disks or disk replacements, CPU overhead again becomes the bottleneck.

1.2.3.3.2.2   DBMS Efficiency

As discussed in Section 8, the DBMS being considered for PDSC appears to consume too much of CPU resources to be appropriate for the application. The DBMS modelled in this study, although instigating approximately 50% more I/Os per request, is much more conservative in its consumption of processor time. The CPU overhead assumed for this DBMS ($\sim$ 13 ms) is an educated guess. Application of a monitor to an operating version of, say SARP V, would be useful in verifying the estimate. SARP V is a multi-level index driven DBMS developed by Bunker-Ramo for the CATIS and is the prototype of our data base modeling. It is reasonable to speculate, however, that a like DBMS could be constructed which would be even more efficient. Combined with multi-threading and/or faster disks or other storage devices, a DBMS which used only, say, 5 ms of CPU resources per I/O could theoretically deliver in the range of 635,000 I/Os per hour (transfer time of 2/3 ms is added to the 5 ms).

The important point being made here is that, through multi-threading and faster storage devices (e.g., high density head-per-track disks as discussed in Volume I, Section 3.0 of [MEAS79], the long-standing data base I/O bottleneck has been or soon will be, resolved. The bottleneck will now be at the CPU and the software within it. If the software can not be appreciably improved by reducing the number of

instructions which must be performed per I/O, then the CPU itself
must be enhanced or augmented by the addition of ancillary special
purpose processors.

1.2.3.3.3 Bottleneck #3 The Central Processor

Two CPU bottlenecks, having different characteristics, are perceived.

1.2.3.3.3.1 The DBMS CPU

As mentioned above, it may be necessary to enhance the speed of the
CPU which is performing data base activities. We feel that this can be
accomplished most cost effectively through the application of tightly
coupled, asynchronous multi-processor minicomputers or even multi-
processor micro-computer configurations specifically designed for data
base operations. The latter can be produced more economically than
general purpose processors and can also be more efficient in performing
their functions. Such a configuration can be called a "Data Base
Machine" and is represented in current technology by the proposed
Gaertner G-471, described in Appendix A.

It is recommended that work continue on developing the G-471 and other
data base machines such as the Mc^2 Hybrid Machine (also described in
Appendix A. Additionally, we feel that it is important that efforts
be undertaken to establish software algorithms for utilizing these
machines.

1.2.3.3.3.2 The Applications CPU

The requirements at the applications CPU are of a different nature
and require a different solution. "Applications" in this context
are considered to be heavily numeric processing functions and matrix
operations. The larger word sizes of mainframe CPUs are considered
more appropriate to this form of activity than are the shorter words

of mini-computers. Characterization of the PDSC system indicates
that these "number crunching" functions are secondary to the data
base and message handling duties it must perform. For this reason,
less expensive minicomputers are considered adequate to perform the
applications node functions for PDSC. An upwards expandable tightly
coupled, asynchronous multi-processor mini is recommended.

The ADCOM mission, however, is much more computationally oriented
and, in our opinion, will be best performed at the applications node
by a long word, fast mainframe. As requirements increase, a main-
frame can be augmented by off-the-shelf special purpose processors
which are tailored to the job, such as vector machines. For future
systems, the next generation, this function will best be performed by
arrays of micro-processors designed to perform specific operations, tied
together following schemes such as that employed in the G-471.

1.3 Approach

The approach is keyed to a phased procedure which is expected to
yield significant early results in system performance enhancement, and
will chart the progression from the present to future mission
fulfillment. The early identification of the technological areas of
greatest potential benefit to the intelligence community, coupled
with the systematic application of technology, will make possible a
far more focused and effective approach to intelligence system
specification, design, implementation and enhancement.

The technical approach may be summarized as follows:

o   Exhaustive characterization of operational and functional
     attributes. The intent was to collect sufficient data to
     identify and substantiate operational problem areas and
     contributing factors.
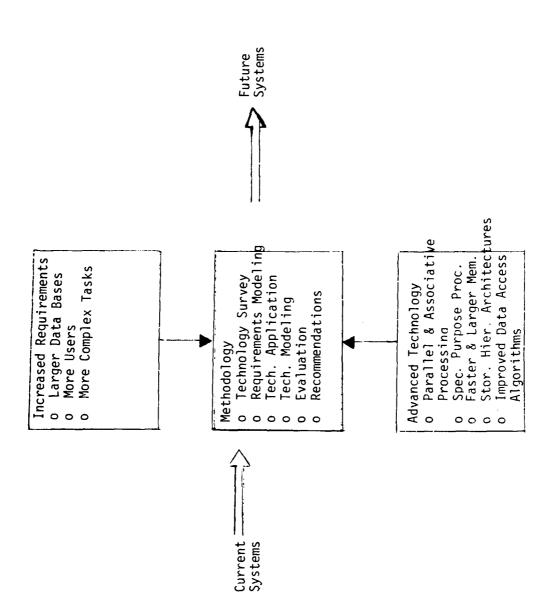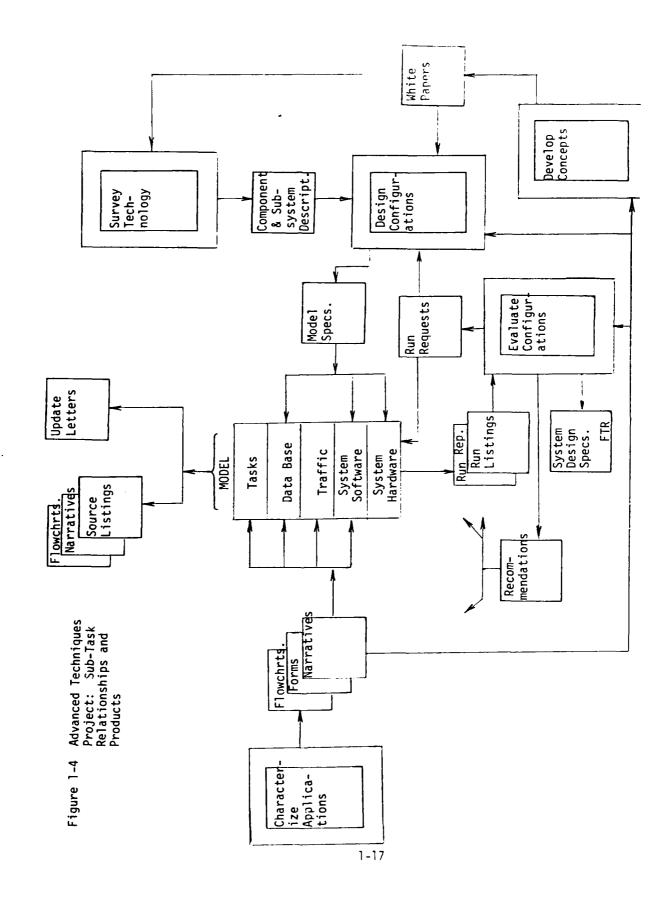
Figure 1-3 Application of Advanced Technology

o   Employment of modeling and simulation techniques for
     demonstrating current or projected timings under various
     system loads; and comparing alternative strategies involving
     insertion of advanced parallel technologies.

o   Assessment of technology to differentiate timeframes (i.e.,
     near-term, future, etc.) for application of potential
     technology. Basis for timeframes is primarily state-of-the-
     art of the applicable technology and development or life
     cycle stage of the intelligence processing system. The
     overriding purpose of this approach is to provide a low
     risk, cost effective, time phased development plan.

Figure 1-3 shows very simply the general procedure in applying
advanced or improved technology to meet increased requirements:

o   Requirements, of course, enter the equation as goals to be met.

o   Available technology enters the equation as building blocks
     for meeting requirements.

o   The methodology performs the synthesis of these elements
     and selects a system design.

o   If a current system exists it may represent a baseline for
     development, thus reducing costs for some candidate
     configurations.

Figure 1-4 shows in somewhat more detail the methodology itself and
the interplay among its elements.

Figure 1-4 Advanced Techniques Project: Sub-Task Relationships and Products

1-17

2.0 THE ENVIRONMENT

Although the history and user requirements of ADCOM and PDSC are
significantly different, both systems (along with other developmental
and operational systems such as SAC/IDHS and AIRES) share demanding
operational requirements to provide increasing numbers of users/
analysts/operators with more reliable, accurate, and correlated
information in shorter periods of time. Having learned a lesson from
the breakdowns in intelligence evaluation and dissemination over the
last two decades, and sensitive to the speed at which similar or more
threatening events unfold in today's world, (e.g., Mideast), the
intelligence community and its customers have been building larger,
faster, and more complex communication networks to carry both timely,
event oriented information (sensor based) and historical data that is
required to assess potential threats and recommend necessary counter-
action. To effectively support local and national level threat
assessment, strategic planning, and sensor management, computer data
bases have swelled with new information and with improved historical
summaries of existing elements; they have become more complex to
support near-real-time analyst queries and to properly correlate the
wealth of data available from the new generation of sensors.

The situation is not unique to strategic, technical, and other national
level intelligence applications; requirements to support tactical
commands are rapidly emerging and are also stretching the technology
base with demands for even faster responsiveness and manipulation of
voluminous target data.

Personal "shoeboxes" and other hardcopy files previously within the
domain of a particular intelligence group are being automated and

disseminated to other elements with a critical need-to-know. As a
result of increased data base size and complexity, more esoteric,
response-consuming computations and higher data base and communications
activity levels, the large computer technology of the 1960s has
rapidly given way to the multiple processor, functionally partitioned,
mini/host technology of the 1970s. Systems such as SAC PACER are
being "expanded" with special dedicated subsystems such as SACWARDANS
and COMPASS PREVIEW, and will be enhanced with a special purpose,
multiple-microcomputer-driven, high bandwidth communications port (MICS).
The ADCOM configuration also shows signs of functional partitioning
through the assignment of separated SOI, SAWS, IDHS and I&W sub-
systems. Technological problems, although temporarily overcome, are
again starting to emerge as further coordination between geographi-
cally dispersed operational elements is mandated and as growing
numbers of user facilities are attached. An excellent example of
the difficulty is provided by the PDSC system concept. In this case,
the existing, standard data base technology would appear to be
inadequate to the task assigned it.

Whereas existing technology will be able to solve many of the problems
in the next few years, there remain a few applications where the
technology has been (or soon will be) pushed to its limits and where
previously anticipated growth requirements may not be met.

Systems of the future will be characterized by much larger and more
volatile data bases, more users, and more complex and data-demanding
functions provided to users. These functions will range from enhanced
graphics support, such as time compressed mobile unit displays, which
requires both increased processor power and copious data retrieval,
to near real time updates of dispersed data base elements.

Coupled with these considerations is an expected great increase in
update activity to support anticipated broadened coverage, and

enhanced resolution and timeliness of information. Figure 2-1
outlines these and other developments which are expected to push
intelligence systems technology well beyond its current capabilities.

MUCH LARGER DATA BASES

     o   MULTI-SENSOR/MULTI-SOURCE CORRELATION

     o   EXPANDED COVERAGE

     o   MORE COMPREHENSIVE, ON-LINE HISTORICAL DATA

HEAVIER TRAFFIC

     o   MORE USERS

     o   HEAVIER DEMAND PER USER

          -   ENHANCED INTELLIGENCE ANALYST AIDS

              --   ENHANCED CORRELATION FUNCTIONS

              --   ENHANCED GRAPHICS SUPPORT

                  --- BAR CHARTS

                  --- CURVES

                  --- COLOR

                  --- MAP OVERLAYS

                  --- 4D

     o   DRAMATICALLY HIGHER UPDATE RATES

REQUIREMENT FOR FASTER RESPONSES

FASTER MOVING EVENTS BEING TRACKED

Figure 2-1   Increased Requirements

3.0 TECHNOLOGY SURVEY

To form a basic "list of materials" to be applied to the problems at hand, a survey of state-of-the-art technology was performed. In some cases this technology was projected or, we hope, improved upon. Working papers were produced and delivered throughout the duration of the project:

o   "Architectural Alternatives in Data Base Management Machinery", October 1978.

o   "Technology Survey (State-of-the-Art)", November 1978.

o   Numerous project memos.

The technology survey is summarized in Appendix A, covering the following items:

o   Storage Technology
-   Solid State Memory Technology
--   Metal-Oxide-Semiconductor Random Access Memory (MOS RAM)
---   Static
---   Dynamic
--   Charged-Coupled Devices (CCDs)
--   Magnetic Bubble Memories (MBMs)
-   Electronic Disks
--   Paging Disks
--   Disk Caches
-   Moving Head Disk Technology
--   3330 Technology
--   Winchester Technology
--   Advanced Technology

- --- CDC 33502
- --- Thin-film heads
- Optical Disks
- Miscellaneous
  - -- Laser-Holography
  - -- Laser Bit Address
  - -- Electron Beam Address
  - -- Tunable Dye Laser
  - -- Magneto-Optic Beam Address
  - -- Amorphous Semi-Conductors
  - -- Josephson Devices
- Storage Hierarchy
  - -- INFOPLEX (MIT)
o Data Base Management Machinery
- Data Streaming
  - -- AFP (OSI)
- Hybrid Data Base Machine (Mc^2)
- Associative Processing
  - -- ACAM (Stanford U.)
  - -- REM (Semionics)
  - -- Micro-APP Chip (Brunel University)
  - -- ALAP (Hughes Aircraft)
  - -- PEPE (BMDATC, U. S. Army)
  - -- HAPPE (Honeywell)
  - -- RAP (Raytheon)
  - -- STARAN (Goodyear)
  - -- OMEN (Sanders Associates)
  - -- SCAT (Sanders Associates)
  - -- EGPP (U. of Erlangen)
  - -- AFP (General Precision)
  - -- CASSM (U. of Florida)

- -- RAP (U. of Toronto)
- -- RARES (U. of Utah)
- -- ECAM (Honeywell)
- -- DBC (Ohio State U.)
- -- DIRECT (U. of Wisconsin)
- -- G-471 (Gaertner)

The survey is reproduced in Appendix A of this report.

4.0 CHARACTERIZATION

4.1 Background

The characterization phase of the I&W Advanced Parallel Processing Study involved analysis of the two subject systems, ADCOM and PDSC. Each system has been subjected to a careful study and the operating characteristics and requirements for each function have been extracted from the preliminary documentation available. Some of the candidate system functions are currently operational in some form and data has been collected concerning these functions. Experience with other intelligence systems has been drawn upon to evaluate and estimate characteristics where hard data is missing.

During the characterization phase of this project a great quantity of documentation was perused, particularly information concerning the PDSC System currently being developed. This documentation provided basic input concerning program sizes, message, query and report flow, and system operating characteristics. The quantity and level of detail obtained during characterization was too voluminous to attempt to model and obtain usable statistics. As a result, a study was performed to determine what level of detail could be handled by the models, then a top-down structured approach was used to sift the data gathered, remove the non-essential elements, mold the basic required activities into building blocks by which the models could be easily constructed, then modified as required. The document entitled "A Method of Selecting Equipment Improvements for Intelligence Data Processing Networks" describes the terminology and techniques which were used to further refine the characterization data into modelable information.

4.2 Building Blocks

To utilize an analytic or simulation model to predict the performance
of different computer systems, it is prerequisite to define the
expected distribution of the workload among the system's components.
One must also describe both the hardware and software resources in
some elementary form. These elementary forms can then be used as
building blocks to define more complex functions which in turn are
looked upon as available resources by user oriented applications
programs and system software. The number of building blocks at each
level of characterization must necessarily be limited to those most
frequently utilized or those which demand significant system resources.
The building blocks selected at each level were chosen from the point
of view of how representative they were with respect to the intelligence
system being characterized.

Scenarios were defined for each of three operational conditions:
"planning", "alert", and "crisis". The basic activities which must be
performed for each operational condition are relatively fixed. The
distribution of intelligence functions within each scenario basically
defines the workload over the length of time during which a particular
operational level exists.

The projected workload is also affected by the capabilities of the
hardware resources on which it is executed. The type of equipment
used influences the behavior of the basic building blocks. Within
the context of predicting the response of an existing computer system,
the hardware descriptions will not change from day to day. However,
when testing candidate system architectures against the three secnarios,
this is the only part of the model which will change.

Building blocks utilized both in the characterization and in the
models have been defined at three levels. Intelligence Functions
(IFs) are functions performed by intelligence staff personnel at different
military headquarters and within different commands. Although the

terminology varies, Intelligence Functions are similar for most intelligence systems. Intelligence Functions are composed of a series of Data Processing Functions (DPFs). These are computer-oriented functions basic to any data processing system. DPFs are comprised of Processes which execute Primitives, the lowest level of cnaracterization used to define CPU instructions and I/Os executed. Figure 4-1 shows the relationship between IFs, DPFs, Processes and Primitives.

4.2.1   Intelligence Functions

An examination was made of several different lists of the functions performed by intelligence staff personnel. Although the terminology varies within different commands, similarity can be detected between these lists. It aids understanding to name several groups of related functions or "functional areas". A list has been prepared that consists of 25 intelligence functions grouped into five functional areas. These functions have been chosen so that they could be used to describe both the ADCOM and the PDSC systems, but it is also possible to describe any intelligence system in these terms. A list of the Intelligence Functions may be found in Figure 4-2.

FUNCTIONAL AREA

Situation Monitoring —

INTELLIGENCE
FUNCTIONS (IFs)

ELINT
PHOTINT
COMINT
Space Object ID —

DATA PROCESSING
FUNCTIONS (DPFs)

Comm Handling
Msg Processing
Applications
Data Base Query

PROCESSES

Data Receipt
Logging & Storage
Data Distribution
Data Retrieval

PRIMITIVES

Allocate Storage
Release Storage
Store Data on Device
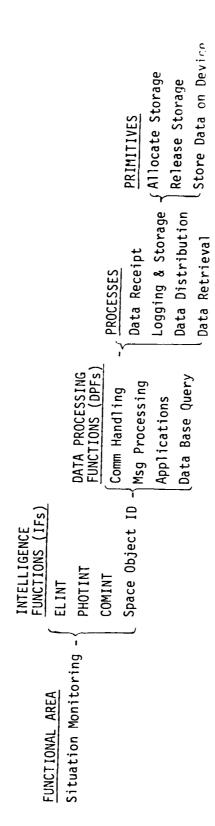
Figure 4-1 Level of Intelligence System Characteristics Modeled

FUNCTION	DESCRIPTION
1. Communications Handling	Management of the receipt,
1.1 Comm Handling In	retrieval, and transmission
1.2 Comm Handling Out	of data via communication lines.
1.3 Comm Handling Retrieval	
2. Message Processing	Management of the processing
2.1 Message Receipt	requirements of discrete sets
2.2 Message Retrieval	of data called messages in
	Intelligence Systems.
3. Applications Programs	Programs (processes) which per-
3.1 Edit AP (PDSC)	form specific operations upon
3.2 Extract AP (PDSC)	data. (These operations are
3.3 Correlator AP (PDSC)	peculiar to the specific system.)
3.4 Auto-ID AP (PDSC)	
3.5 Auto-Call AP (PDSC)	
3.6 Area Search AP (PDSC)/3.2 Analysis (ADCOM)	
3.7 Computational AP (PDSC)/3.1 Computational (ADCOM)	
3.9 Graphics APs (PDSC)/3.5A Graphics, 3.5B FFT (ADCOM)	
3.10 Format Conversion AP (PDSC)	
3.11 Duplicate Data Check AP (PDSC)	
4. User (Analyst) Support	Management of general services
4.1 Pending Action/Notice Q Update	available to the users (analysts)
& Review	of the system.
4.2 Work File Processing	
4.3 Hold Queue Processing	
4.4 DB Command Mode	
4.5 Analyst-to-Analyst Communication	
4.6 Message Generation	
4.9 Graphics Interface	
5. Data Base Update Services	Management of file maintenance
5.1 Data Base Update	processing for add, change and
5.2 Data Base Update Review	deletion of data from the data
	base (A level above DBMS).
6. Data Base Query	Management of processing of
6.1 Simple Query	user query requests for infor-
6.2 Intermediate Query	mation from the Data Base (A
6.3 Complex Query	level above DBMS).

Figure 4-2 Data Processing Functions Modeled Page 1 of 2

FUNCTION	DESCRIPTION
7. Report Generator	Provision of standard report
7.1 Simple RPG	formatting and dissemination
7.2 Intermediate RPG	processing.
7.3 Complex RPG	
7.4 Super Complex RPG	
8. Output Device Management	Management of various forms
8.1 Simple Output Device Management	of output devices and media
8.2 Intermediate Output Device Mgt.	(printers, plotters, graphics
8.3 Complex Output Device Management	display terminals, teletypes,
8.4 Super Complex Output Device Mgt.	CRTs, etc.).

Figure 4-2 Data Processing Functions Modeled   Page 2 of 2

5.0 ANALYSIS TECHNIQUES

5.1 Analysis Tools

Two major modeling approaches were used to predict intelligence system performance: analytic and simulation. An analytic solution seeks equations relating the desired performance parameters to the modeled workload and the equipment configuration. A simulation model is used when the equations defining this relationship are either unknown or insoluble; the model generates a possible behavior sequence of the system and measures the desired performance quantities in that sequence. An analytic model is generally far less expensive to use than a simulation model of the same system. Even though it is not as accurate, it is able to provide "ballpark" estimates of required hardware and expected performance for a given workload. The analytic model described in this report has been used to provide estimates for required hardware configuration, workload redistribution and validation of characterization workload modeling and simulation modeling results.

It must be remembered that these tools provide only approximate solutions to the problem of determining computer system performance characteristics.

The language used for the simulation model is ECSS II (Extendable Computer System Simulator II) as implemented on the Honeywell 6180 computer at RADC. The analytic model is programmed in interactive FORTRAN on the same computer. It utilizes standard queueing theory algorithms for central-server computer architectures.

A more detailed discussion of the models can be found in Section 5 of [MEAS79].

6.0 SUMMARY OF ADCOM STUDY

6.1 System Description - ADCOM

The ADCOM Intelligence Center (ADIC) system is a multi-computer
system which provides automated support to the ADCOM Intelligence
Functions.

Figure 6-1 shows the basic ADCOM three node configuration. The four
node configuration (Figure 6-2) was used as baseline to model
estimated increased demands on the ADCOM system.

The primary functions of the ADIC are:

    o   to provide the CINCAD with near realtime intelligence

    o   to receive and to disseminate strategic and tactical
         warning information

    o   to assess current air, space and missile threats to
         the North American continent

    o   to assess the residual threat

In addition, the ADIC serves as a part of the Worldwide Indications
and Warning System. Two of the (major) important automatic functions
performed by the ADIC computer system are initial target signature
analyses for Space Object Identification and reduction of sensor
data received by the system.
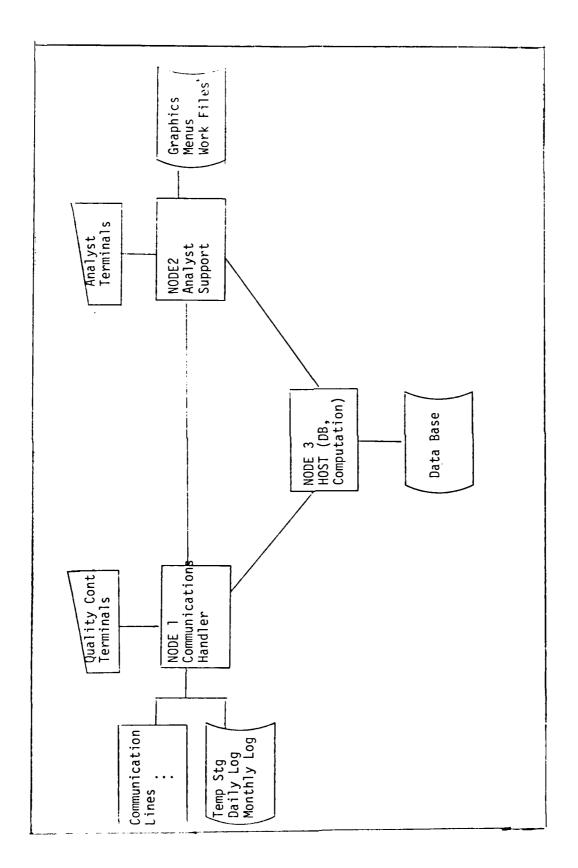
Figure 6-1 ADCOM - 3 Node Original Configuration
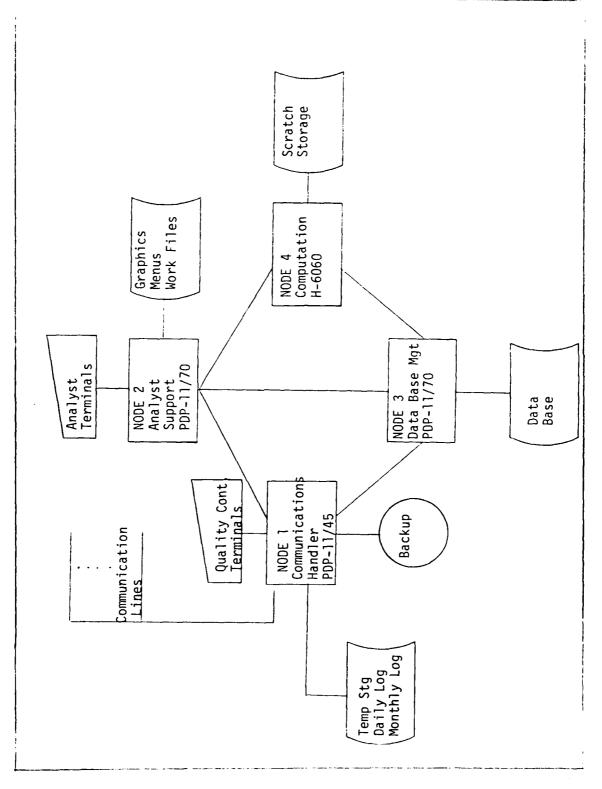
Figure 6.2 ADCOM - 4 Node Baseline Configuration

6.2 Investigation Summary and Recommendations

As described in Volume I, Section 5, of [MEAS79], the ADCOM
Intelligence System was modeled using both simulation and analytic
models. Analytic modeling results served to fine-tune the
simulation model and to cross-check simulation results for the
baseline configuration. Because the accuracy of an analytic model
decreases with heavier loading, only the simulation model was used
to determine results of heavier loading and modified software and
hardware enhancements to respond to the increased system demands.

6.2.1 ADCOM HOST Modeling Results

Figure 6-3 contains a summary of pertinent statistics resulting from
exercise of the ADCOM models. The major findings which are
represented in this figure are:

- o   Need for more computational power. As characterized, the
      computation oriented applications of the ADCOM environment,
      if all get done, will consume approximately 120% of a HIS
      6060 central processing unit (.55 MIPS). A CPU capable of
      1 MIPS, could perform the initial workload adequately with
      comfortable near term expansion capability.

- o   Importance of Multi-Threaded Data Management. The greatest
      single improvement in data access efficiency observed was
      that of multi-threading the DBMS to permit overlap of disk
      seek and transfer times if they occurred on different
      drives.

A fair comparison of the different DBMS configurations can be made by
comparing the CPU utilization at the Application Node. Since the
application programs must wait for returns from their requests to the

6-4

Book No.	Configuration	Load	% Utilization				Approx. % Load Perf.	Approx. # I/Os at DB Node	Response Time for 20 I/O Query (secs)		
			DB Node CPU	Appl Node CPU	Disk System	Special Enhanc.			Average	Minimum	Maximum
4	3 Node, Single-Thread, Baseline	1	-	69.0	71.0	-	54	49,868	26.0	2.4	95.8
5	4 Node, Single-Thread	1	32	67.0	62.0	-	56	55,999	17.8	1.0	62.9
7	4 Node, Single-Thread, Disk Cache	1	32	78.0	22.0	3.3	65	59,073	2.4	.5	7.8
6	4 Node, Multi-Thread	1	35	94.0	9.0	-	78	65,320	1.3	1.0	2.7
8	4 Node, Multi-Thread, Disk Cache	1	35	93.0	8.0	4.0	78	68,199	1.8	.8	6.9
9	4 Node, Multi-Thread, Index on CC Disk	1	34	93.0	1.7	1.2	78	64,153	.9	.3	7.8
10	4 Node, Single-Thread	3	32	5.4	71.2	-	5	64,496	502.0	170.0	908.9
11	4 Node, Single-Thread, Index on CC Disk	3	56	37.0	.7	2.7	31	129,903	70.0	1.9	185.2
12	4 Node, Single-Thread, Disk Cache	3	41	13.0	31.0	4.7	11	86,148	190.1	84.2	269.9
13	4 Node, Multi-Thread	3	79	82.0	26.0	-	68	188,016	4.9	3.5	10.1
14	4 Node, Multi-Thread, Index on CC Disk	3	74	68.0	3.7	3.3	57	177,516	11.9	2.6	24.8
15	4 Node, Multi-Thread, Disk Cache	3	75	74.0	8.1	9.9	62	178,350	12.8	5.1	26.1
16	4 Node, Multi-Thread, Super Disk	3	81	83.0	2.0	-	69	195,999	4.7	.9	22.5

Figure 6-3 ADCOM Simulation Summary

Data Base Node, their execution is an indirect measure of the effect-
iveness of the data access process. The independent nature of
queries arriving from the Analyst Support Node somewhat clouds the
picture if an attempt is made to measure data access effectiveness
by observing the number of I/Os performed. With this in mind, a
comparison of the several configurations will show that whereas
adding a disk cache to a single-threaded system improves performance
approximately 16% (78%/67%), a multi-threaded DBMS, with no additional
hardware, improves performance by 40% (94%/67%).

Addition of a disk cache or a "CCDisk" for indices to a multi-threaded
system actually degrades performance slightly (1%).

This can be explained in part by the single-threaded nature of the
devices and in part, perhaps, by the random arrival distributions of
the model. At any rate, a multi-threaded disk system, giving, possibly,
a 10 ms access time if an average of 4 disks are concurrently operating,
will be limited by the CPU overhead associated with each I/O (13 ms
in the model). If effective disk access is overlapped 100% with
processor overhead, the effective access time is 13 ms.

The observed simulated response times support this analysis in that the
average and maximum response times for the disk cache configuration
(1.8 and 6.9 respectively) are greater than those for the straight
multi-threaded system. The minimum response time for the disk cache can be
expected to be shorter since, when a request arrives at an empty or near
empty queue, access time is shorter for each I/O. The same holds true
for the "CCDisk" System as to minimum response time. The fact that
"CCDisk" average and maximum response times were shorter than those
for the basic multi-threaded system, probably reflects the small
size of the sample (seven queries over one hour).

6.2.2 Analysis

The ADCOM model provides figures which can be used to specify
hardware requirements. It indicates, first, that the 11/45 at the
Communications Node (26% utilized) and the 11/70 at the Analyst
Support Node (14% utilized) are adequate to process the expected load
easily and should easily accommodate anticipated increases. For
purposes of brevity, statistics for these two nodes were not included
in the table. They may be found in Appendix I of Volume II of
[MEAS79]. The Honeywell 6060 (rated in this study at .55 MIPS) at
the Host Node, however, will be inadequate for even the initial load.

6.2.2.1 Option 1 - 3 Node System, Large Mainframe HOST

It would probably prove sufficient to replace the H6060 with a 1 MIPS
mainframe for the initial load. This judgement is based on the
following calculations using statistics from Figure 6-3:

$$
\begin{aligned}
\text{Data Base:} \quad & .28 \text{ MIPS @ 35\% doing 78\% of requested work} \\
= \; & .126 \text{ MIPS @ 100\% doing 100\% of requested work} \\[1em]
\text{Applications:} \quad & .55 \text{ MIPS @ 94\% doing 78\% of requested work} \\
= \; & .663 \text{ MIPS @ 100\% doing 100\% of requested work} \\
& \therefore \; \text{\# MIPS required} = .126 + .663 = .789 \\
& \therefore \; \text{A 1 MIPS mainframe could perform the work at} \\
& \quad \sim 79\% \text{ utilization.}
\end{aligned}
$$

For the intermediate term, the 1 MIPS mainframe would prove inadequate,
based on a 50% increase in workload. At least a 2 MIPS machine should be
contemplated, although somewhat less would suffice, e.g., 1.5 MIPS for a
utilization of 80%.

6.2.2.2   Option 2 - 3 Node System Phased to 4 Node System

The 1 MIPS mainframe could be used for initial loads and a PDP-11/74* with two CPUs (effective MIPS ~ .56) could later be installed to assume data base processing duties.   In this case, under 150% of initial load, the 1 MIPS mainframe would be 99% utilized (66% X 1.5). To alleviate this situation, one of three courses of action can be followed:

o   The 1 MIPS machine chosen for the initial configuration should be one that is easily enhanced (modularly expandable) to at least 1.5 MIPS, or,

o   A 1.5 MIPS or better mainframe should be chosen for the initial configuration, or,

o   A special purpose array (or vector) processing machine can be added to the configuration when required, to perform the heavy matrix (FFT, etc.) operations.

Please note that Option 1 requires a 1.5 MIPS mainframe, whereas Option 2 requires approximately the same plus an 11/74 equivalent. The reasons for presenting, and preferring, Option 2 follow:

o   The Option 2 mainframe is less utilized, thus improving response time and retreating from the dangerously high utilization estimate of 80%.

o   Option 2 is a 4 node configuration, separating data base activities from numerical/application activities.

---

*Common memory with from one to four central processing units which operate in parallel, asynchronous fashion. The PDP-11/74 is not the only such machine which might be considered. It is chosen here as a representative of the genre for ease of exposition and for its compatibility with the other DEC equipment in the system.

6.2.2.2.1 Separation of Data Base and Numerical/Application Activities

For reasons of expandability in the long term it is deemed advisable to separate, from the very beginning, the two major and dissimilar activities of the system: data base management, and numerical/ scientific applications. If this step is taken now, at moderate cost, incidentally, not only will system efficiency be enhanced for the near and intermediate terms, the system will be in a flexible stance for long term expansion to meet as yet un-quantified but almost certain-to-be greatly magnified requirements. If data access and/or update requirements outrun the capabilities of even a four processor 11/74 type machine, the entire Data Base node can be replaced with something more powerful with minimum system upheaval. A data base machine such as those surveyed in Volume I of [MEAS79] (the Hybrid Machine, for instance) could be easily introduced to the system as the replacement for the "11/74".

If computations/applications outstrip the capabilities of the mainframe, special purpose processors can be added to the Application Node in a manner transparent to the system. Alternatively, a more powerful mainframe could be easily introduced.

6.2.2.3 Option 3 - 4 Node System

Similar to Option 2, this option calls for a four node system. The difference is that it begins as a four node system. This option will be more expensive in the initial stages since it requires both a mainframe and an 11/74 type machine from the beginning. In the intermediate term, however, life cycle costs considered, Option 3 will be more economical. By beginning with a one processor "11/74", the transition to heavier loads can be accommodated by adding one or more processors. Option 2, which begins with only the mainframe,

would require new programming effort to move the data base services
to the new machine when it is added. Also, an additional shake-down
period would be experienced while adjusting to the new installation.
The preferred development schedule is, then:

Phase 1 (Initial)

Data Base Node   -   "11/74" with one CPU

Application Node   -   1.5 MIPS or better mainframe, or
                                   a 1 MIPS mainframe which is easily
                                   expandable to 1.5 MIPS (preferably
                                   more).

Phase 2 (+ 5 - 6 Years)

Data Base Node   -   Add another CPU to the "11/74"

Application Node   -   Unchanged

Phase 3

Data Base Node   -   If required, replace "11/74" with
                                   a data base machine. Addition of
                                   up to two more CPUs to the "11/74"
                                   may suffice.

Application Node   -   As required, add special purpose
                                   vector type machines, or replace
                                   mainframe with more powerful
                                   machine.

Other modifications to the original, baseline, configuration are
summarized as follows:

o   Communications line speeds have been adjusted upwards to
       handle the projected baseline traffic

o   Quality control terminals at the Communication Handler node
     have been increased from 1 to 4

o   Number of analyst terminals required to handle baseline
     loading at Node 2 vary from 27 to 32. Original
     characterization indicated 30.

Each of these areas will need to be modified upward to include any
necessary multiplexers/controllers as terminals and communications lines
increase.

6.2.3 Evaluation of the ADCOM Simulation Model

The ADCOM simulation model is only as accurate as the data used to
build it. The accuracy was most apparent when trying different loads
and configurations. The results were sometimes surprising and at times
it seemed that they must be incorrect, but they proved to be correct, if
not what was expected, when they were closely examined. An example
of this occurred when comparing s Single-Threaded DBMS, Multi-Threaded
DBMS, Single-Threaded DBMS with cache, and a Multi-Threaded DBMS with
cache. The Single-Threaded DBMS with cache was faster than the single-
threaded alone and the multi-threaded was even faster yet. It would
seem that the Multi-Threaded DBMS with cache would be even faster, but it
was not. It appeared that the Multi-Threaded DBMS with cache was not
as efficient as the Multi-Threaded DBMS because, although 8 disks could
be used concurrently, there was only one cache, and it was, in effect,
single-threaded and thus a limiting factor.

The simulation model also proved very flexible when a single
statement was added to the DBMS portion causing three times as much
disk and processor activity to be simulated for each DBMS request.
When this modification was used with several different configurations
involving multi-threading and other DBMS enhancements, it appeared
that the 11/70 processor was a major limiting factor.

The flexibility of the model was also shown with the various user
statistics that were collected and printed. During the development
many different statistics were collected. Most of these were
discontinued due to both the volume of printed data and the fact
that, although many statistics were useful for developing the model,
they were redundant and often not useful to the final model.

6-12

The accuracy and utility of the results from the ECSS II ADCOM model depends on numerous factors, some of the more important are:

o   Detail modeled

o   Accuracy of processes of jobs being modeled

o   Correctness of hardware description

o   Correctness of message distribution and arrival rates

The detail modeled in the ADCOM simulation has a large influence on the accuracy and usefulness of the results. Obviously a more detailed model will produce more detailed and correct results, but this is only true if all of the extra details used are accurate and if they produce noticeable change. The gross details used in constructing the model are relatively easy to obtain and verify. Such things as message arrival rate over given communications lines, the average length of the messages, and the Intelligence Functions that receive the messages have been tabulated and are simple to use in the model. They are also easy to change, should the tabulated rates be shown to be incorrect or the rates changed to simulate projected rates. Other details such as instructions executed per process and number of I/Os are not as well known and not as easy to change in the simulation, due to the nature of ECSS II and the way the model was implemented.

More detail also results in a larger and slower running model. An example of this would be a process that executes 12,000,000 instructions and does 300 DBMS reads and writes. The 12,000,000 instructions should not be executed all at one time since this would tie the simulated processor up and prevent other simulated jobs from running for unrealistically large blocks of time. A better method would be to execute 100,000 instructions 120 times or 10,000 instructions 1200 times. The latter is probably more realistic but will be much less efficient to simulate. The DBMS activity should also be broken up

into 300 separate calls intermixed with the instructions, rather than one call for 300 reads.

Better input for the simulation model could be obtained by using hardware and software monitors on existing intelligence systems, since this would give actual counts of instructions and I/OS performed.

The accuracy of hardware descriptions and capabilities is very important, but two errors were found after it was too late to make any more simulation runs. The speed of the unibus was given as 4,000,000 bytes/second. This was taken from DEC publications, but appears to be incorrect for the model since this figure includes processor and memory control signals. A more nearly correct figure for simulation purposes would be 2,000,000 bytes/sec. While this is a 100% difference, the effect on the model should be slight since the fastest device being modeled has a transfer rate of less than 1,000,000 bytes/sec. The error will only show up with multiple concurrent transfers.

Another error appeared in the way the Multi-Thread DBMS was implemented. The disk seek time and rotational delay time were lumped together; on a Multi-Threaded DBMS they should be separate since seek time on one disk can overlap any operation on another disk but rotational delay and transfer time cannot overlap. The way the model was written will allow overlapping as if each disk had a separate controller.

The model was correct in that it would not allow any overlapping
of operations of one disk with itself.

This inaccuracy caused the effectiveness of the Multi-Threaded DBMS
to be overstated. While the utilization figures and process time for
the DBMS are off by a relatively small amount, it is not thought
the error is enough to change the judgement that the processor is
the limiting factor for the faster DBMS.

The tabulated message distribution and arrival rates were used in the
model, but at first the printed results were not what was desired.
This was due to the method of simulating message arrivals and the
small sample being used. It is expected that unless an infinitely
large sample is used, the distribution may not be what is desired.
If a given message was to occur 36 times per hour the message was
triggered with an exponential distribution centered about 100 seconds.
Due to the randomness of the exponential distribution and the
relatively small number of samples the desired number of messages was
not often not reached. This was most evident when only 6-12
messages/hour were expected. This situation was corrected by multiplying
the interarrival times by adjustment factors until the printed results
agreed with the tabulated rates. This is effective because the
random numbers used for the exponential distribution are not truly
random, but are pseudo random and are repeatable, (i.e., each
simulation run will produce identical numbers). This is only valid
for runs of the same time period, from 1400 to 5000 seconds.

6.3 Recommended ADCOM Configurations

6.3.1 Baseline Configuration

As discussed in Section 6.2.2.3 of this volume, we recommend Option 3-
a baseline 4-node system utilizing a "PDP-11/74" type machine for the
DBMS Host and a mainframe Applications Host. Figure 6-4 depicts the
Phase 1 overview of the 4-node ADCOM System.
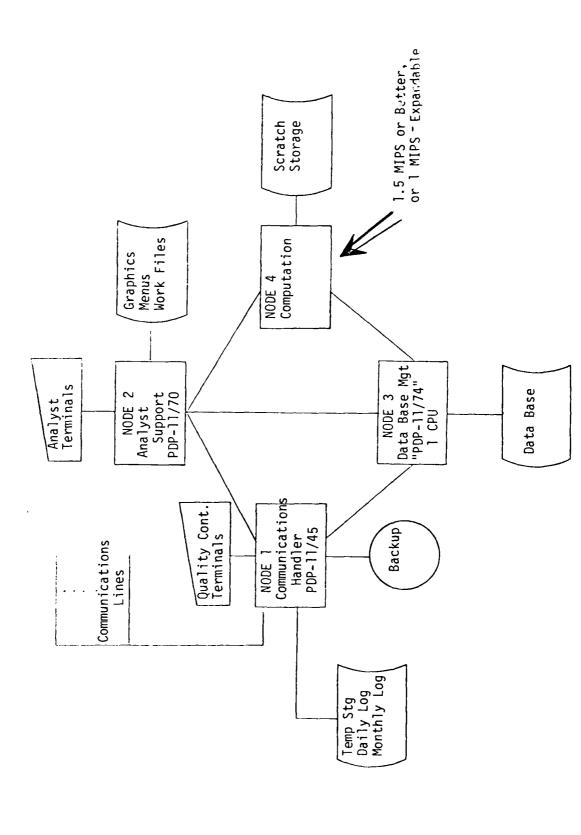
6.3.2 Intermediate Configuration

For Phase 2 Intermediate expansion, within the 5 to 6 - year time period,
we foresee a need to add another CPU to the "11/74" DBMS Host
Configuration. Figure 6-5 gives the Phase 2 ADCOM Overview.

6.3.3 Long-Term Configuration
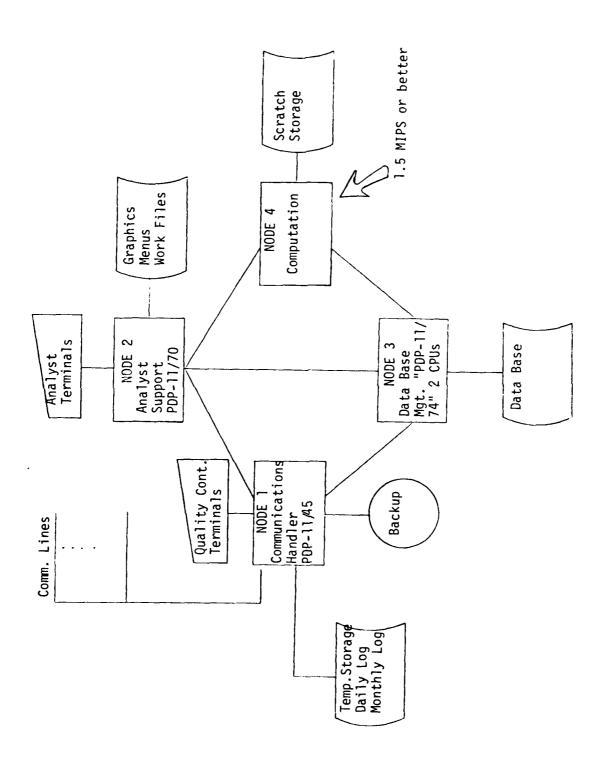
Phase 3, a projection of 8-10 years from baseline will require charges
to both the Applications and DBMS hosts:

o   DBMS - Add 2 more "11/74" CPUs to Phase 2 configuration or
     replace with a data base machine

o   Applications - add special purpose vector type machines to
     Phase 2 configuration, or replace mainframe with more powerful
     machine.

6-16
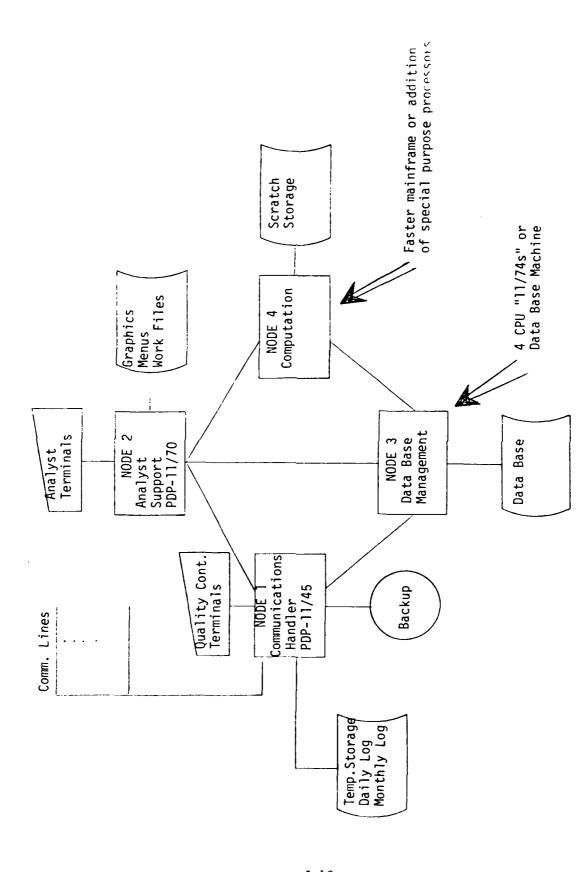
Figure 6-4 ADCOM System Overview: Phase 1 - Initial

Figure 6-5 ADCOM System Overview - Phase 2 - Intermediate (+5-6 Years)

Figure 6·6 ADCOM System Overview – Phase 3 – Long Term (8-10 Years)
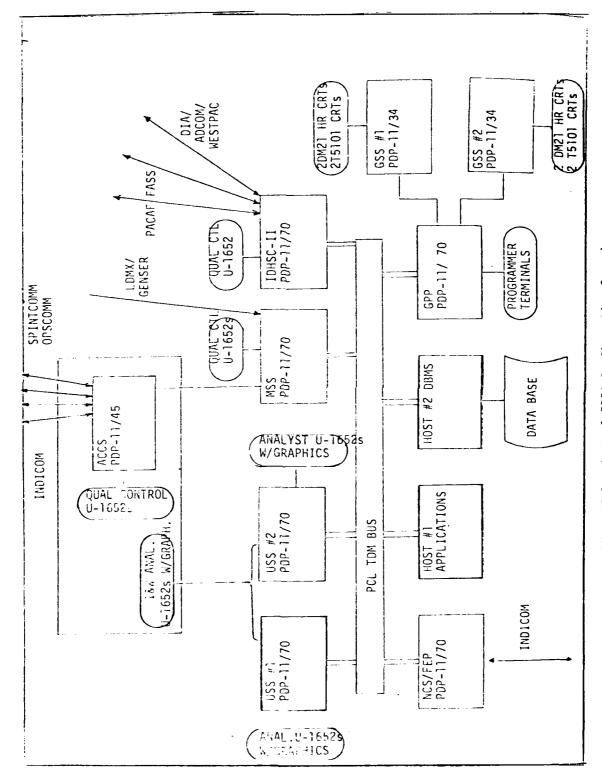
7.0 SUMMARY OF PDSC STUDY

7.1 System Description - PDSC

The Pacific Command (PACOM) Data System Center (PDSC) is an intelligence
network of geographically separated computer centers in support of the
PACOM mission. The mission involves data collection and intelligence
dissemination concerning an area of approximately 96 million square
miles. Essential functional areas of responsibility for the PDSC are:

o   To assess/estimate the near and long-term developments in the
       Pacific theater and provide intelligence in support of
       decisions on the locations, kinds and size of force dispositions.

o   To provide warning, of impending trouble in support of decisions
       to change the readiness posture of forces.

o   Development of targets and weapons recommendations for
       contingency and general war planning and associated defense
       analyses.

o   Indepth and timely intelligence support of component air,
       land, and naval forces with orders of battle, target data,
       logistic intelligence.

The PDSC brings the intelligence analyst's data handling capabilities
in line with the capabilities of special information-collection and
communications systems. On-line interactive access to All-Source and
collateral Data Bases being fed by near-real-time external sources
enable analysts to support the PDSC Mission. Figure 7-1 shows an overview
of the central PDSC Configuration.

Figure 7-1   Central PDSC Configuration Overview

7.2 Investigation Summary & Recommendations

This section outlines the steps that were taken to evaluate the efficiency and applicability of proposed PDSC Intelligence System Develop.

7.2.1 PDSC Characterization

Section 4 of Volume I of [MEAS79] describes the methodology used in the characterization work done for ADCOM and PDSC. Also explained are the workload modeling techniques, an extension of characterization. Details of PDSC Characterization have been included as appendices to Volume III of [MEAS79]. Appendices include:

   o   PDSC Intelligence Function Flow
   o   PDSC Workload Modeling Charts
   o   Data Processing Function Charts
   o   Process Flowcharts
   o   Instruction and I/O Rates
   o   Charts and diagrams

Note that these charts and diagrams are basic building blocks for any intelligence system. We hope that the data will be reviewed and Mc^2 will be given any adjustments/corrections to the information as characterized. We feel powerful tools have been designed that could, with fine tuning, be used to measure both current and proposed intelligence systems performance.

Results of PDSC Characterization and Workload Modeling for the Baseline, Crisis - Peak Hour Load have been given in Appendix II of Volume III of [MEAS79]. Figure 7·2 shows results of workload projections: Crisis Baseline Hourly + 4 years. Loading was computed by multiplying transactions for all Intelligence Functions except IF #5 - ELINT & EOB by 1.5. IF #5 - ELINT and EOB transactions

PDSC HOST1 & HOST2 Crisis Hourly Loading Summary - Totals, Baseline + 4 Years (Cont'd)

Total HOST1 & HOST2 DBMS Instructions Executed: 1,220,658,633
Total HOST1 & HOST2 DBMS I/Os Executed: 459,758
Instructions Executed per DBMS I/O: 2,655

Total HOST1 & HOST2 Instructions Executed: 3,024,226,927
Total HOST1 & HOST2 i/Os Executed: 618,378
Instructions Executed per I/O   4,891

Notes

1. The term "Rate" used for instructions and I/Os indicates the number of instructions and I/Os executed per single DPF occurrence.

2. The "Occur" column under HOST1 indicates the number of times a DPF is initiated at HOST1, although DBMS processing at HOST2 may also occur in support of the DPF. The "Occur" column under HOST2 shows the DPFs which are initiated and completed at HOST2.

3. Instruction and I/O Rates for DPF No. 3.2 - Extract Application:

   AP Instructions: (Occur*3200)+(1960*(n-Occur))÷Occur=Average Instructions per Occurrence

   AP I/Os: (2*n)÷Occur=Average I/Os per Occurrence

   DBMS I/Os: (n*5.7)÷Occur=Average DBMS I/Os per Occurrence

   DBMS Instructions: Avg DBMS I/Os per Occur*2655=Average DBMS Instructions

   The values of n by Intelligence Function are: IF#1, n=57; IF#6, n=29; IF#7, n=5; IF#8, n=420; IF#9, n=3,855.

Figure 7-2   PDSC HOST1 & HOST2 Loading Summary Totals - Baseline + 4 Years

7.2.2 Evaluation of PDSC Modeling

During the ADCOM modeling effort on this project it soon became
apparent that a simulation model of the PDSC would be impractical.
The ADCOM Model consumed excessive computer system resources, both
in core requirements ($\sim$ 100K words) and in processing time ($\sim$ 1.5 hours
of CPU time per run). It was obvious that the PDSC, being a
considerably larger system than the ADCOM, and having more diverse
functions, would prove unmanageable as a simulation model given the
resources of this project. The decision was taken, then, to rely on
the extensive and detailed characterization of the PDSC exhibited in
the appendices to Vol. III of [MEAS79] and upon the Analytic Model, which,
although it requires considerable preparation of input (viz. the
characterization summaries), is thrifty with computer resources.

Some discussion is required concerning the results of the analytic
modeling of PDSC in this volume.

The DBMS Host CPU utilization figures for both IOC and IOC+4 years (81.6%
and 129.2%, respectively) are considered valid projections within the
context of traffic volume assumed. They indicate that more than the power
of a single 11/70 will be required to process currently anticipated loads.

The Applications Host CPU utilization figures, also for both IOC and IOC+4
years (80.4% and 170.9%, respectively) are more open to question since they
are based on assumptions concerning the logic and efficiency of applications
programs. Such routines are by nature less predictable than system soft-
ware and packaged DBMSs because their functional specifications and
quality of implementation can vary widely from system to system. The

figures obtained, however, indicate a situation similar to that of the
DBMS Host. The indicated utilization figures call for more computing
power than has been configured. It should be noted that these figures
include graphics processing, a significant contributor to the load. It
has not yet been determined, to the best of our knowledge, that graphics
processing will be performed on the Applications Host.

The Message Support Subsystem (MSS) CPU utilization of 6.5% for IOC seems
rather low and is suspect. Considerable effort has been devoted to
identify any shortcomings in the characterization with only minor modifi-
cations resulting. When compared to NMIC observed loading of 10%,
the figure is at least in the approximate range. At such low utilization
the accuracy is a moot point, since no increase in processing power is
necessary even for double the traffic. There is some opinion in the field,
however, that the PDSC MSS will be much more heavily utilized than the
NMIC MSS, due, perhaps, to more sophisticated text processing functions
now available. This area could probably benefit from further investigation.

Other nodes exhibit such low utilization that no problem is indicated with
the current configurations. Should new information come to light con-
cerning the loading at these nodes, they can be reconsidered.

We feel that a simulation model of the PDSC would be a valuable tool for
on-going system development and tuning. Whereas, unfortunately, the
characteristics of ECSS II, in itself the most efficient simulation
language we found available, made construction of this model impossible
given the resources of the current project, we feel that such a model
might be possible at a future data.

During the construction of the ADCOM model certain causes of the excessive
core and CPU requirements were discerned by inference. One such cause
was long queues. When the system was capable of processing the load

expeditiously, thus keeping queues short, less core was required to run
the simulation. There exists, to our knowledge, no documentation or study
dealing with ECSS II efficiency considerations. A worthwhile project would
be a short study employing controlled experiments with ECSS II to identify
and quantify its resource consuming characteristics to provide guidelines
for creating more efficient models. With this information in hand, it
might very well be possible to create a streamlined, efficient PDSC
simulation model.

7.2.3 Analysis

Based on the results of the characterization, modeling, and evaluation of
the proposed PDSC configuration and loading estimates, it has become
apparent that the only indicated problems are in the Host DBMS and
Applications areas. It is our understanding that the Host is to be divided
into functionally separate processors, one for data base processing, and
one for applications processing. We concur heartily with this decision.
With a functionally distributed Host, enhancements specifically aimed
at either data base or applications requirements can be cost effectively
applied with minimum disruption of the system. If, for instance, in the
farther future, it is found that data base processing requirements
outstrip even the most powerful of processors available, a data base
machine (e.g., the Hybrid Machine described in Appendix A) can be
substituted for the DBMS Host without disturbing the applications
services. Likewise a new Applications Host can be installed, if necessary,
without disturbing data base services.

It is recommended that at both the DBMS Host and the Applications Host,
a PDP 11/74-type machine be installed. This machine is configurable with
one memory and from one to four 11/70-type central processing units which
operate in parallel, asynchronous fashion. Processors may be added as
needed (up to 4) with little effort.

7-7

The PDP-11/74 is not the only such machine which might be considered. It is chosen here as a representative of the genre for ease of exposition and for its compatibility with the other DEC equipment in the system.

The recommended phased development plan for the PDSC DBMS Host and Applications Host is:

Phase 1 - (Initial)

DBMS Host - "11/74" with 2 CPUs
Applications Host - "11/74" with 2 CPUs

Phase 2 - (Intermediate - +4 years)

DBMS Host - add one CPU to "11/74"
Applications Host - add two CPUs to "11/74"

Phase 3 - (Long Term - +8 - 10 years)
DBMS Host - add fourth CPU to "11/74", or, if developing requirments warrant, replace "11/74" with a data base machine.

Applications Host - add special purpose (e.g., vector) processors.
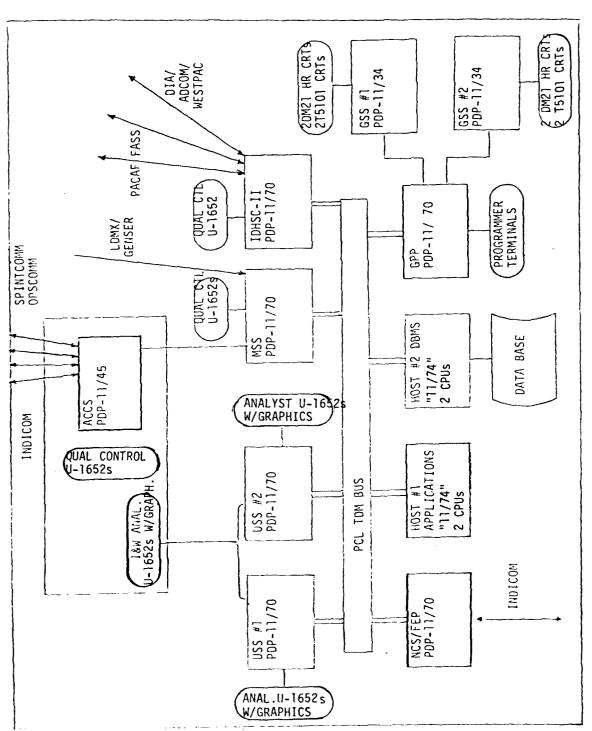
7.3 Recommended PDSC Configurations

7.3.1 Baseline Configuration

Although the CPUs planned for PDSC IOC will handle the initial
workload, both Applications and DBMS HOST CPU Utilizations - IOC,
Crisis Peak Hour, are uncomfortably high. Utilization above 70%
usually results in increased response times - an unacceptable
condition in a cirsis or "conflict" situation. Therefore Phase 1
recommendations shown in Figure 7-3, are to use a "PDP-11/74"
with 2 CPUs for each HOST node. Modeling indicates that communi-
cations line speeds must be increased to handle crisis traffic. The
quality control U-1652s, if redistributed over the three "Interface
Nodes" ACCS, MSS, and IDHSC-II, will suffice for baseline loading.
Analyst U-1652s probably need to be increased at the two USS nodes
from 80 to 95 (reference PDSC Model Book #4, Appendix I of [MEAS79].
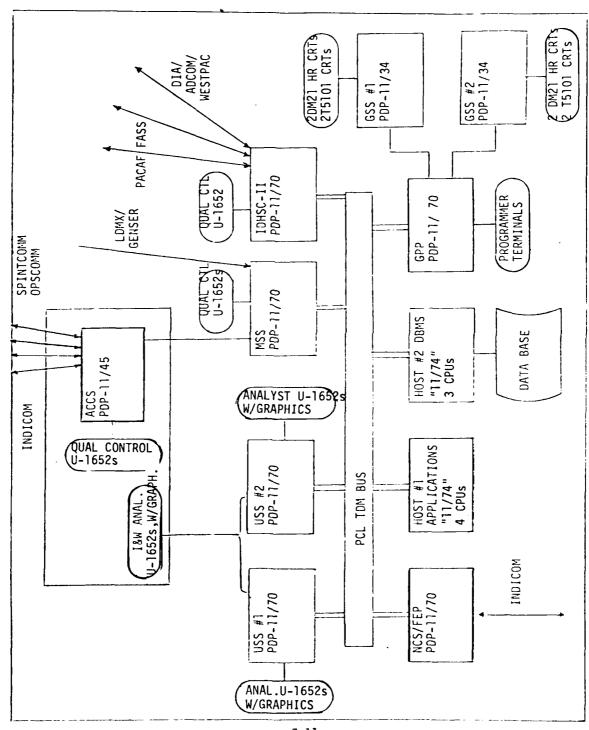
7.3.2 Intermediate Configuration

Results of analytic modeling of the PDSC Baseline + 4 years, a projected
overall central PDSC load of 1.55% of IOC, shows requirement for double
the CPU capacity at the Host nodes, with no room for growth. Figure
7-4 depicts the proposed configuration for Phase 2 - 4 years from IOC.
Increase in loading of 20% will require increased communications line
speeds and at least 1 more quality control terminal at both ACCS
and MSS nodes. Analyst terminals at the USS computer center will
need to be increased from 95 to 120, with a corresponding increase
in multiplexers.

Figure 7-3   Central PDSC Overview: Phase 1 - Initial

Figure 7-4   Central PDSC Overview: Phase 2 – Intermediate (+4 years)

7.3.3 Long Term Configuration

If the projected rate of growth continues or increases, the HOST configuration depicted for Phase 2 will not handle the loading. By IOC + 8 to 10 years, communications lines again will need to be increased; quality control terminals for the three interface nodes (ACCS, MSS, ACCS) will increase from 9 to 11; Analyst terminals of USS will be increased from 120 to 150 with a corresponding increase in multiplexers. Change from the PDP-11/45 to an 11/70 at ACCS would provide an improvement in response time due to the faster CPU and a superior busing scheme. The USS nodes may require an "11/74" type configuration to handle 150 terminals. Figure 7-5 depicts the proposed long term configuration - Baseline +8 to 10 years.

If a Data Base Machine is required at the DBMS Host, the Hybrid Machine, as described in Appendix A is a prime candidate. New development required to construct this machine would not be of an unreasonable magnitude. It would consist of the design and implementation of new software for the three new elements: the Control Processor, the Statistics Processor (if implemented), and the Streaming Processor. The Random Processor could probably use existing, off the shelf, data management software. Also required would be the development of a disk system having two sets of movable heads, driven by separate controllers, or high capacity, head per track disks involving technology such as thin film heads (see Paragraph 3.2.3.3 of Appendix A). Depending on the implementation, a new disk for the Random Processor would have to be purchased as well as bulk memory for the Random Processor and the Statistics Processor.
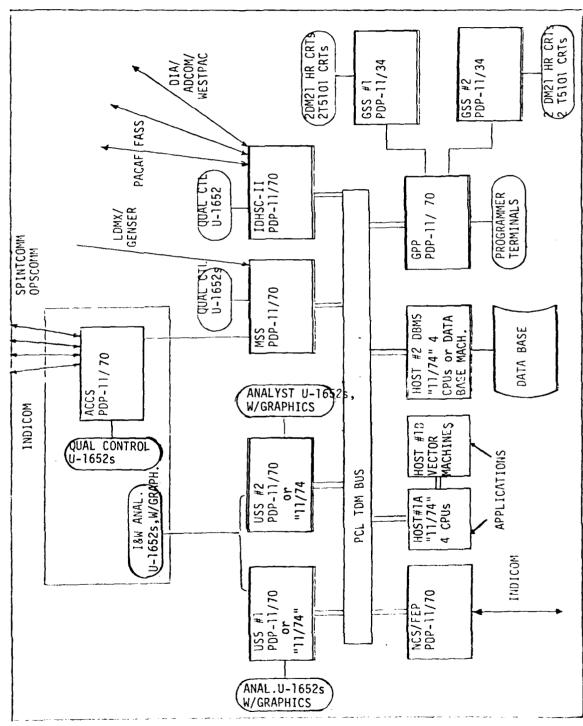
Figure 7-5   Central PDSC Overview: Phase 3 - Long Term (+8-10 Years)

7-13

It is possible that the "11/74" with 3 CPUs, serving as the Phase 2
DBMS Host could perform the Hybrid Machine functions of the Control
Processor, the Random Processor, and the Statistics Processor. The
Streaming Processor and associated peripherals must be specified
and developed separately.

8.0 A SOFTWARE CONSIDERATION

The foregoing analysis and recommendations assume a multi-level, index driven DBMS. The following discussion is included to bring to notice a clear indication that DBMS-11, a CODASYL based DBMS for the PDP-11 series, as currently implemented, will impose a severe burden on the system due to extraordinarily high CPU utilization requirements. The discussion is presented in an ADCOM context because the ADCOM situation viz a viz this issue is somewhat less complex than the PDSC situation and is felt to be more appropriate to this summary report. For a treatment of the PDSC implications, refer to Volume III of [MEAS79].

8.1 DBMS-11 Performance Analysis

The CPU utilization estimates employed to model DBMS-11 are based on the MITRE document, PACOM Data Systems Center (PDSC) Prototype Data Base Performance Analysis, September 1978. Average CPU time per I/O is derived based on the measured CPU time to perform simple and intermediate data base queries reported in the MITRE document. Under DBMS-11, an average of 3.8 reads per single query is assumed. This value includes 1.8 reads associated with an average FIND CALC operation and two reads of non-contiguous detail records. An intermediate query is assumed to be equivalent to 10 simple queries; i.e., 38 reads. The approach used in this data base performance analysis was to employ the "worst case" MITRE test results because of the fact that the prototype data base modeled in their study was much smaller than the expected ADCOM data base.

The method employed in this analysis to arrive at an "average" CPU time per I/O uses the "last 5/non-contiguous COBOL" timings presented for simple and intermediate queries presented in Tables III and IV of the MITRE report. We add to these observed values, the timings for subschema binding (35 ms), record binding (8 ms) and journalizing (52 ms) for a total overhead of 95 ms. The following Figure 8-1 summarizes these times.

8-1

Query Type	Measured Timing (ms)	Overhead (ms)	Total I/O & CPU	CPU Util.	Total CPU (ms)	CPU Time per I/O (ms)	No. CPU Inst.	No. of Occurrences
Simple Query	235	95	330	.85	280.5	73.8	20668	988
Intermediate Query	3810	95	3905	.84	3280.2	86.3	24170	321

This DBMS-11 I/O is being performed on a PDP-11/70 operating at .28 MIPS. If 85% of 330 ms is being used for a simple query requiring 3.8 I/Os and 84% of 3905 ms is being used for intermediate queries requiring 38 I/Os, the average number of CPU instructions executed per DBMS-11 I/O is:

Simple: $(.330)(.85)(.28 \text{ MIPS})/3.8 = 20668$ instructions

Intermediate: $(3.905)(.84)(.28 \text{ MIPS})/38 = 24170$ instructions

Average: $\frac{988}{1309}(20668) + \frac{321}{1309}(24170) = 21527$ instructions

Figure 8-1   Conversion of DBMS-11 Observed Request
Timings to CPU Instructions per I/O

The ADCOM DBMS Node is performing approximately 70,000 data base I/Os per hour. If we assume a DBMS-11 requiring 21527 CPU instructions per I/O where the number of DBMS I/Os is reduced by 1/3*, we execute:

$$(21527 \text{ instr/IO})(2/3)(70,000 \text{ instr.}) = 1,004,593,300 \text{ instr./hour}$$
$$= 279,053 \text{ instr./second}$$

If a single PDP-11/70 operates at 280,000 instructions per second, then the lower bound on CPU utilization due to DBMS alone is 100%, equivalent to a full PDP-11/70. This figure compares to a CPU utilization of approximately 35% for the multi-level, index driven DBMS as modelled.

8.2 Suggestions

If this analysis is correct, DBMS-11 is unsuitable for the systems examined in this project. It is true that much of the time consumed is in areas which could be improved (e.g., context switching), but the MITRE document indicates (we have not been able to verify) that certain operations are invoked as a result of a request that are central to the structure of DBMS-11 and could not be avoided (e.g., binding and journaling operations). That these operations could be sufficiently streamlined to make DBMS-11 responsive to the application at hand is open to question.

The time given in the MITRE report for journaling (52 ms per request) seems excessive. It could probably be disabled in DBMS-11 and replaced with a system journaling utility of a different form (e.g., a simple recording of the unparsed request which should take no more than a few hundred microseconds).

---

*A multi-level, index driven DBMS is assumed in the model. Such a DBMS is assumed to perform an average of 5.7 I/Os per simple query, as opposed to 3.8 for a CODASYL DBMS.

The binding operations, as reported, take such a large amount of CPU resources (43 ms combined) that one is led to suspect that a translation from original schema and subschema declarations is taking place and that a dynamic binding service is being provided. It would be worthwhile examining the feasibility of replacing this dynamic binding approach with a static binding approach. Some query flexibility would be lost but a study might show that this degree of flexibility is not needed.

It is suspected, based on previous investigations of SARP III and IV, that SARP V is also performing dynamic binding operations. If this is so, the same recommendations made for DBMS-11 hold for SARP V. The DBMS modeled on this project, while SARP-like, was not performing dynamic binding and was therefore much more efficient than SARP V is suspected of being.

9.0 REFERENCES

[MEAS79]   System/Subsystem Specifications (3 Volumes), Mc^2,
            September, 1979.

[MEAS78]   "A Method of Selecting Equipment Improvements for
            Intelligence Data Processing Networks", Mc^2 Working
            Paper, November 1978.

[MITR78]   PACOM Data Systems Center (PDSC) Prototype Data Base
            Performance Analysis, Mitre Corp., September 1978.

APPENDIX A

TECHNOLOGY OVERVIEW

3.0 TECHNOLOGY OVERVIEW

3.1 Introduction

Computer technology has come a long way since the days of the original
ENIAC project, however, no single form of memory has been found to
satisfy all of the storage requirements of computer systems. There
are fast, expensive internal memories for calculations and slower, less
expensive ones for peripheral storage, consequently there is a gap
between them in terms of speed and cost. Computer developers and users
have attempted to compensate for the difference in speed between internal
and peripheral storage. Multiprogramming and virtual storage concepts,
among others, have been invented solely to make up for the speed disparity.
Figure 3.1-1 illustrates the trade-off in access times and memory
capacities between available technologies for main memory and for secondary
storage.

We will at first discuss storage technologies that will change the above
described picture, and eventually discuss their implications to computer
architectures. ⋅ ⋅⋅ ⋅ ⋅⋅ ⋅⋅ data base management machinery.

Total Memory Capacity (bits)

10^{13}
10^{12}
10^{11}
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
10^{5}
10^{4}
10^{3}
10^{2}
10

Bipolar RAM

MOS RAM

Fixed Head Disk

Moving Head Disk

Mass Storage System

Magnetic Tape

10^{-9} 10^{-8} 10^{-7} 10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1} 1 10
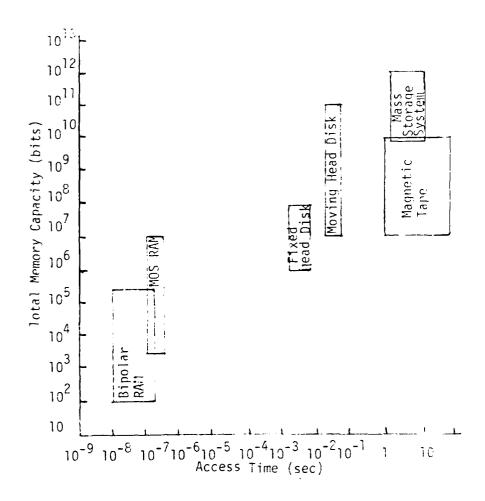
Access Time (sec)

Figure 3.1-1 Memory Capacity vs. Access Time Tradeoff of Current Memory
            Technology

A-2

3.2 Storage Technology

3.2.1 Introduction

New mass storage technologies using holography, lasers, optical disks, magneto-optics, and amorphous semi-conductors are currently in various stages of development and testing.

Developments in this direction are high technology intensive and quite often involve technical breakthroughs. High research and development costs of any of the above systems will in the short run still result in expensive systems having a small production series.

An alternative to this situation could be a more modular, less involved technology. "Electronic disks" have been suggested to realize this objective. The term "electronic disk" refers to electronic storage devices (as opposed to electromechanical or rotating devices) which possess memory capacities and cost-per-bit-of-storage which compete with conventional storage devices. In addition, those small scale storage systems can be used as an intermediate memory between main and archival memory. Many technologies now being investigated have the potential of being used for electronic disks. These include:

o   Charge-coupled devices (CCD)

o   Magnetic bubbles

Initially we will therefore discuss considerations on the lower end of the storage scale.

3.2.1 Solid State Memory Technology

To date, MOS and Bipolar RAM dominates the mainframe primary memory market.
Charged-coupled devices (CCDs) and magnetic domain bubble memory devices (BMs)
are competing to be the memory gap filler. All have dramatically improved in
the last few years in performance/price with respect to bit density/chip, access
time and cost. This increase in performance/price has made these devices
competitive with secondary storage devices such as fixed head disks or
small moving head disks.

The principle areas of use for these new devices are:

    o   Large slow access main memory

    o   Fixed head disk replacements

    o   Disk CACHE mechanisms

    o   Full disk replacements

We will take a closer look at these new devices in the following
sections.

3.2.1.1 MOS RAM

MOS RAM is Random Access Memory which is characterized by the ability to access any storage location within one access time of typically 150-1500 ns. The primary use of RAMs is main memories where the Random Access nature is essential.

There are two types of MOS RAM:

    (a) STATIC

    (b) DYNAMIC

(a) STATIC RAMs require more circuit elements/bit on the chip and thus are not available in as high a density as dynamic RAMs. They are, however, considerably easier to design with. As the name static implies, once the data is stored, it stays there permanently (until loss of power).

STATIC RAMs are available in 1K, 4K, and 16K versions. The 64K versions are not likely to be available until a year or two after the 64K versions of the dynamic RAMs become available.

(b) DYNAMIC RAMs store the bits as charge. This charge leaks off with time and thus, DYNAMIC RAMs must be periodically refreshed or they will lose the stored information. This refreshing must take place approximately every few milliseconds and results in about 1% unavailability.

Presently, DYNAMIC RAMs are available in 1K, 4K, 16K and 64K versions. Larger than 64K versions (256K) will not probably appear in the near future (before 3-5 years).

3.2.1.2 CCDs

CCDs store binary data as charges in long circular shift registers.
As with DYNAMIC RAMs, the charge leaks and must be periodically
refreshed. With each shift, one bit is read, and one bit is
regenerated and thus the system designed need not be concerned with
refreshing as in DYNAMIC RAMs.

The CCDs, however, being shift registers, do have an access time that is
much larger than RAMs. This access time is on the order of 1 ms or
approximately 1000 times that of a RAM. After the access, however, the
transfer rate is comparable to RAM (1/2 to 5 MBytes/sec).

From the above consideration, CCDs have application where either

   o   A Pattern of Access exists (e.g., TV Raster Display), or

   o   Block transfers are the primary use (e.g., disk paging)

CCDs and RAMs suffer from volatility. On loss of power, they lose the
information stored.

While few products exist which use CCDs, as their cost comes down,
they begin to compete with fixed head disks in paging schemes.

3.2.1.3 MBMs

Magnetic Bubble Memories store binary data as small magnetic domains. They are typically organized similar to CCDs and could be used in similar products. The most obvious difference between the MBMs and other devices is their non-volatility. They don't lose their memory with loss of power. In large memory schemes such as disk replacements, this is an important feature.

The other comparable features of MBMs are access time and transfer rate. These times are approximately 5 millisec and 50 Kilobits/sec, respectively. Comparing these with CCDs we see that the CCD has a much shorter access time (1/2 millisec) and much higher transfer rate (5 million bits/sec). The MBMs, however, do come with higher bit densities (100 K, 256 K bits/chip) than the CCDs.

Bubble technology at this point in time can deliver 1 Mbit modules (Intel. Magnetic 7110). New techniques that should be able to improve on this performance are:

o   Cross Hatch Memory (IBM, San Jose)/Current Access Method
     (Bell Laboratories)
     Coils implementing the rotating magnetic field in a magnetic
     bubble device have kept down improvements possible for bubble
     devices. This field-access technology can be replaced by current-
     access technology. The technique relies on two conductive
     sheets punctured with rows of elongated holes. Every hole
     on one sheet overlaps with the ends of two holes on the
     other sheet. Current flowing in the sheets will create the
     fields necessary to move the bubbles around.

o   Contiguous disk devices (IBM, Yorktown Heights)

The name of this technique comes from the shape of the patterns
that guide the propagation of the bubbles. Contiguous disks
do not need the critical interpattern spacing resolution that
are required by chevron-shaped patterns common to today's
bubble chips. The resulting density would make it possible
in principle to store 25 Mbits on an area that was used for
TI's original 92 kbit device. An additional benefit from this
technology is the possibility of stacking layers - therefore
making realistic a three-dimensional memory array with densities
of hundreds of megabits.

o   Wall-encoding (IBM, San Jose)

Currently bubbles must be separated by four to five
diameters, in order to prevent the bubbles from interacting
with each other. One way to overcome this restriction is to
use bubbles of two different types. Wall-encoding, that is
inducement of changes in the magnetic structure of the wall
region that separates each bubble from the surrounding
material, does just that. One type of structure represents
a "one" and the other a "zero".

Combined with contiguous disk technology this wall-encoding technique
could result, theoretically in a 16 million bits/centimeter density
device.

Much work has also been done researching the proper layout of the
bubble device to minimize access time of the data. Topology considera-
tions relative to major and minor loops have also lead to the introduction
of the cache concept in the bubble module domain.

Ultimately considerations have been given to the place of bubble devices
in data base management systems. The concept has been expressed of
having in one module part of a relational data base as well as a bubble
logic search processor   [CHAN78].

3.2.2 Electronic Disks

With the lowering cost of solid state memories, the idea of using them for disk replacements has become attractive. The principle characteristics of these devices are:

o   Low access time

o   Medium cost

o   No moving parts

o   High reliability

The areas where these devices will be applied in large size configurations are:

o   Add on, large, slow access main memory

o   Replace fixed head disk

o   Main memory paging schemes, caches

There is one basic driving force for using solid state memories as disk replacements. There is the much lower access time (1/2 ms vs. 40 ms) of these devices. The need for lower access time in disks can easily be seen. Typical main memories have access times on the order of 1 us. The fastest fixed head disks have average access times of 8 ms. This is a performance factor of 8000:1. This usually results in systems that are I/O bound, with the processor mostly idle.

The hardware Survey [MEAS78b] details some existing bulk-core, CCD and IBM products.

The technology used for disk replacement will largely influence the characteristics of the resulting product.

3.2.2.1 Disk Replacements Based on CCDs

CCDs are the most, inexpensive (.1 cent/bit), available in high
densities (64K), and have very high data transfer rates 5 Mbit/sec.
However, due to their volatility, these devices can not serve as
large (300 MByte) disk replacements. Thus CCD based disk replacements
are basically limited to paging schemes or disk caches.

3.2.2.1.1 Paging Disks

Paging disks are usually fixed head disks or drums. They are used to
cache blocks of main memory in mainframe computers.

A computer may have only a one megabyte main memory, but several
concurrent programs each using 2 megabytes of memory. The programs
actually reside on the paging devices, with only a part in the main
memory. When a part of the program that is not in memory needs to execute,
the paging device swaps the least used page in main memory with the
desired page on disk. During this swapping, a different program is
allowed to run.

CCDs make ideal paging devices because they have much lower access times
than disks or drums and can be configured in sizes comparable to these
devices at competitive prices. The only drawback to CCDs, their
volatility, is not a problem with a paging device because the system
looks at the paging device as though it were main memory, not as a permanent
file structure such as a data file.

3.2.2.1.2 Fixed Head Disks

The major application of fixed head disks is its use in paging schemes. However, certain applications may require temporary high speed access such as manipulating large arrays of data. CCDs would be applicable here for the same reasons as in the paging schemes.

3.2.2.1.3 Disk Caches

A disk cache is a large buffer that is used to store the most recently referenced disk blocks. A typical size buffer would be approximately 1/8 to 20 megabytes, for a disk size of 80-300 megabytes. The cache device sits between the I/O controller and the disk drives. The I/O handler in the main computer does not see this cache device. This has the advantage of being software transparent.

Upon requests from the I/O handler for particular disk blocks, the cache buffer is automatically checked to see if the requested block is in the buffer. If it is, a request to the disk drive need not be made and the response time will be ~1/2 ms instead of ~50 ms. If the block is not in the buffer, disk I/O will take place and the requested block will be delivered to the I/O handler and will replace the least used block in the cache buffer. Further requests for this block will not cause a disk I/O.

For a disk cache to be successful, there are two important criteria that must be satisfied. They are:

  (1) High hit rate

  (2) Low variance in hit rate with varying application

The hit rate is percentage of disk requests that are in the cache buffer. Unless the hit rate is high (>50%), little performance gain will be

achieved. Even if a high hit rate is achieved, the variance of the hit
rate must be low for varying applications.

To illustrate the importance of a high hit rate, a numeric example will
be used.

> Given: Access times for disk I/O and cache I/O as 50 ms and 1/2 ms
> respectively, the following hit rates yield the given average
> access times:

Hit Rate	Avg. Access Time
10%	45 ms
25%	38 ms
50%	25 ms
75%	13 ms
90%	5 ms

We can see that with less than 50% hit rate, performance is improved very
little (<50%). A hit rate of 50% halves the access time and a hit rate of
75% quarters access time.

The importance of little variance of hit rate with varying application can
be seen with the following example:

> Given the above access times, and two applications, the first
> application has an 80% hit rate while the 2nd has only 60% hit
> rate. The average access times for those two applications are:

Application	Avg. Access Time
(1) 80% hit rate	10 ms
(2) 60% hit rate	20 ms

we see that varying the hit rate by only 25% doubled the access time. Thus at high hit rates, small changes in hit rate cause large changes in access time. The hit rate is a function of the cache buffer size compared to the disk size, the application, and the algorithm for determining what blocks are kept in the cache. The hit rate could be made 100% by merely making the size of the buffer the size of the disk. This, however, is the extreme and if this was required the device would be a disk replacement instead of a cache.

The importance of the application can't be overlooked. In large batch type of applications most I/Os take place in a limited area of one or two disk files. High hit rates are possible with small cache buffers by merely keeping entire tracks of a requested block in the cache. In DBMS applications, index files are likely to be requested more often than the data files, and are much smaller than the data files. Consequently, high hit rates may be achieved with small caches by keeping a large percentage of the index files in the cache.

3.2.2.2   Disk Replacements Based on MBMs

Magnetic bubble memories are relatively new with only two current manufacturers and 3 bubble chips commercially available. In the next few years they promise to see wide spread application.

MBMs are the highest density chip of any of the solid state memories available (92K, 256K, 1M). Their price is still high (comparable to RAM) for use as disk replacements but the price should be lower in the near future. MBMs currently available have medium access times (4-10 ms) and transfer rates

(50-100 Kbits/sec). Paralleling them for higher throughput can be done but requires extensive support circuity.

The prime virtue of the MBM is their non-volatility. A complete disk replacement could be implemented without any moving parts. This feature along with the extremely high density makes them a prime candidate for medium size disk replacements.

MBMs have too slow an access time to be used as either fixed head disks or disk caches. (Comparable to present day fixed head disks).

MBMs as medium size disks offer two major benefits. Access time comparable to fixed head disks, and high reliability due to their solid state nature.

The current price of MBMs is about .1 - .2 cents/bit at the chip level. To be competitive with medium size disks (40-100 MBytes), this price would have to drop to about .01 - .03 cents/bit. This is a price factor of from 3 to 10 and will be achieved soon (1 - 2 years).

3.2.2.3 Disk Replacements Based on MOS RAM

RAM competes with CCD memories directly. Any product that can be implemented in CCDs could be implemented in RAMs with no penalty in speed.

Access time of RAM is ~1uS with transfer rates from 1/2 to 5 mbytes/sec. Interfacing is easy and versatile. Virtually any configuration can be achieved easily.

A-14

The only factor that leads one to choose RAMs over CCDs is
availability. The per chip cost of CCD and RAM is about the same.

3 2.2.4 Disk Replacements Based upon Bulk Core

Bulk core is a new product based on the old technology of core memories.
It is currently competitive with any other solid state memory for disk
replacements.

It has all of the desirable characteristics of a solid state memory.
These include:

o   Cost Competitive with other technologies

o   Fast access time (<3 usec)

o   High transfer rate up to 5 MByte/sec

o   Non-volatile

The above characteristics either equal or beat any other form of solid
state memory. Why then, is there little consideration to these devices
in applications where CCDs and MBMs have been applied? The answer is in
the prospects for the future.

MBMs, CCDs, and RAMs all show promise for future increases in density and
price reductions. Core memories show little promise for future price
reductions. Thus, while they are competitive today, it is not likely that
they will reduce in price.

3.2.3 Moving Head Disk Technology

In comparing the currently available moving head disk drives, the most significant identifying characteristic is capacity. Currently available capacities range from 20 to 635 megabytes (MBytes).

The other parameters which characterize all of these drives are:

o   25 - 40 ms average access time (voice coil positioning)

o   4000 - 6000 BPI recording density

o   200 - 400 tracks per inch lateral recording density

o   800 - 1200 kilobytes/sec data transfer rate

o   8 - 12 ms avg. latency (3600 - 2400 RPM)

The small capacity disks (20 - 50 megabytes) are distinguished from the medium capacity disks (50 - 100 megabytes) primarily by the number of platters (sides). The same technology is used with more platters and read/write reads to achieve a higher capacity.

The large capacity disks (100-300 megabytes) are characterized by an increase in capacity through increasing the bit density (BPI) and track density (TPI), along with the use of more efficient recording codes. The increase in recording density is accomplished by decreasing the head height above the recording surface.

There are two categories of large capacity disks with the head height (hence recording density) being the determining quality. These two categories are the "3330" and "Winchester" technology referring to the IBM representatives of these categories.

3.2.3.1 3330 Type Technology

The 3330 type disks physically resemble the medium capacity disks
with removable multiplatter disk packs. It has a recording head
height of only 45 microinches which allows the increase in recording
density from 2000 BPI and 200 TPI to 4000 BPI and 400 TPI. Disk
capacities of up to 300 megabytes based upon this technology are
commercially available.

3.2.3.2 Winchester Technology

The Winchester technology is basically an extension to the "3330" concept
of decreasing the head height. In these type drives the head height is
decreased to 20 microinches. The resulting increase in recording density
is 6000 BPI and 400+ TPI.

The price paid for this increase in density, however, is the necessity
of sealing the heads and media into a dust proof module. As a result,
most Winchester type disks have fixed (i.e., non-removable) platters.
There are a few, however, with removable cartridge type (3340) media where
the heads and platters are sealed in one module. Disk capacities of
up to 600 megabytes are available using this technology.

3.2.3.3 Advanced Disk Technology

A summary of the currently available moving head disk drives is presented
as a part of the Hardware Survey [MEAS78b].

Two disk drives deserve special attention because they are the only
representatives of a particular technology and they are "state of the art"
drives. They are:

    (1) CDC 33502
    (2) AMPEX DM-PTD

The CDC 33502 is one of the largest capacity disk to date and is approximate
twice the size of its competitors. The disk drive uses Winchester
technology and other manufacturers are likely to follow with similar products.

The AMPEX DM-PTD is unique in that it reads 9 of the 19 read/write heads
simultaneously for a data transfer rate of 10 MBytes/sec. In other
respects, the disk is a 300 megabytes 3330 type drive. This drive would
be ideal for operations that require extremely high throughput of
serialized data.

Disks can not be easily dismissed from consideration as the principle
storage medium for the large data base systems of the future.
Although their imminent demise has been forecast for many years, con-
tinuing advances in technological areas applicable to disks have pushed
the capabilities of disk systems consistently ahead of both requirements
and competitors. The anxiety felt by the general computer using com-
munity concerning future applicability of disk storage devices has been
based on the fear that they are too small and too slow and that their
potential has been almost reached; current disk technology provides access
times of approximately 25-40 milliseconds and storage volumes in the
range of approximately .6 gigabytes per spindle (~ 5 gigabytes per eight
spindle system).

Breakthroughs in both read/write head technology and surface coating
processes, however, have made possible a near-term dramatic enhancement
of disks. Thin film technology has made it possible to create finer
resolution heads which are at the same time much lighter and less expensive
to produce. The enhanced resolution of these heads enables a much denser
packing of bits per track as well as tracks per surface. With denser
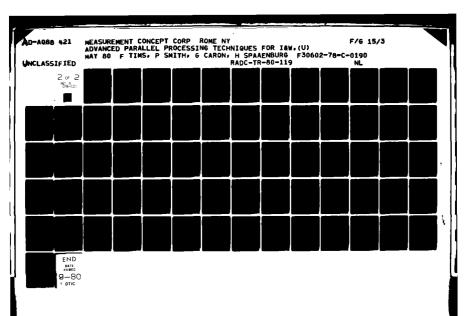
packing and closer tracks, it becomes necessary to position the head closer to the surface. New surface coating processes will soon make this possible by providing a smoother surface. Industry sources predict that these improvements will in the near future, make possible a disk having a capacity of 5 gigabytes. An eight spindle system would contain 40 gigabytes.

Coupled with the increased storage capability is the fact that the thin-film heads, being much lighter, can be more numerous. Head per track devices are now conceivable with the storage capacity of movable head devices. This has heretofore been unachievable due to the resolution of the hand-made ferrite heads. Fixed head (head per track) disks currently store in the range of 10 megabytes.

Another expectation is that rotation speeds will be tripled within the next few years. The disk storage medium we can now contemplate by the year 1985 will have the following characteristics:

o   40 gigabytes of storage for an eight spindle system - made possible by denser recording and finer resolution read/write heads.

o   Average access time of less than 3 ms - made possible by lighter heads (head per track) and faster revolution.

o   Lower cost per bit than current disk systems - made possible by lower head fabrication expense and elimination of arm mechanics (head per track).

For data hungry applications the new read/write head technology combined with the LSI processor capabilities, offers outstanding improvements in

throughput, providing dramatic response improvements and opening the door
to the implementation of much more complex functions which have until now
been unrealizable.

With a modicum of intelligence installed in or close to the read/write
heads to perform compare operations and an appropriately designed
architecture of microprocessors to receive and process data, a data base
machine could be built, based probably on the relational model of data,
which would deliver data to thousands of users concurrently (terminal
queries or running routines) with an average access time to a given
element of data of approximately 1.4 milliseconds, processing the entire
data base in approximately 2.7 milliseconds.

While it would certainly be improvident to halt pursuit of other
technologies such as EBAM with its trillion byte storage possibilities
and access time below 10 microseconds, it is clearly too early to
abandon disks.

3.2.4 Optical Disks

Digital Optical Disk recording is a development that has its basis
in video recorders research. Capacitive and laser techniques have been
investigated in regard to video recording. Currently the attractiveness
of digital data storage realizations using the same techniques
has surfaced products of that kind. In the basic setup
a laser beam is focused onto a spinning disk and then burns in
pit marks representing the data. A disk is read by illuminating the
disk with a laser beam at a lower intensity and observing the subsequent
reflections from the marks on the disk.

The following advantages have been found:

  o   Small spot size of laser beam, makes efficient use of recording
       material. It is possible to store in the order of 10^{11} bits
       of data on a 12 inch disk.

  o   Access times compare favorably to those of magnetic disks.

  o   Cost per disk range from between $15 to $150.

  o   Long range life can be expected.

Problems still exist in the area of focus errors, tracking errors and
signal-to-noise ratios.

Two devices are generally known; those manufactured by RCA and Philips.
A third one, by Hitachi, has been announced. Each of these systems uses
different techniques for laser, recording medium and tracking. The laser
is designed to be high powered; powerful enough to create the pits in a
short time. In addition one would like a relatively long-life laser.

A-21

The recording medium construction allows for the laser limitations by localizing the burn-in effect. The disk also provides for a layering to minimize the effects of dust, etc.

Optical disks, at this point in time, are usable mainly in a read-only mode, although provisions can be made for updates at later times. Therefore the advantages of optical disk will become evermore clear when progress is made in reusable and erasable recording materials. And, as already indicated, lasers have to be developed with a long life. and high power performance.

..2.4.1 RCA Corporation

The optical disk developed by RCA is capable of storing 10^{11} bits of data on a 12-inch disk. The recording medium is a trilayer disk, written onto by an Argon laser. Data recording and readout can be accomplished at rates of up to 50 Mbits per second on a single channel basis.

3.2.4.2 Philips Laboratories

In Philips' DRAW (Direct Read After Write) optical disk system 10^{10} bits can be stored per side of a 30 cm disk. Typical of the Philips approach is a pre-grooved air-sandwich recording medium. In addition to the grooves (simplifying tracking), sector addresses have been pre-marked on the disk. Data rates of at least 50 Mbits/second are anticipated. Storage capacities of up to 10^{15} bits can be accomplished in a moderately sized "juke box" arrangement.

Philips, in cooperation with the Magiec Division of Magnavox, will be shipping prototypes of the dual-sided disk recorder by the fourth quarter of 1980. Drive price with controller will be $150,000. Disks will be priced at $150 each.

3.2.4.3 Hitachi Ltd.

The optical recorder made by Hitachi differs from the RCA and Philips approach in the thin-film material used in the recording medium. It is Hitachi's opinion that the materials selected will result in a disk with a longer-life and less sensitivity to handling. A prototype 500 Mbyte data recorder has been produced. Both the Philips and Hitachi recorder use a read-after-write error detection technique; they both make extensive use of error-correcting coding.

3.2.4.4 Drexler Technology Corp.

Disks as used by RCA, Philips and Hitachi can still be prone to substrate errors in the recording medium. Drexler's approach to these kind of errors is to pre-scan the disks and mark those tracks that cannot be used reliably. The pre-formatted disks thus created will, of course, be more expensive in material but cheaper in usage.

3.2.5 Miscellaneous Emerging Technologies

3.2.5.1 Laser-Holography Storage Devices

Use of holographic techniques to optically store information on archival material is the technology used by Harris Intertype to construct the HRMR (Human-Readable, Machine Readable) storage device, built under contract to Rome Air Development Center. The HRMR was designed to record in three modes:

o   Machine readable binary data

o   Graphic data with its equivalent machine readable form

o   Purely graphic (human readable)

To remain comparable to other technologies, we shall emphasize only the machine-readable recording mode, although the other two modes have definite utility.

The HRMR was designed to provide large quantities of archival quality read-only storage. Fast access times were not a priority as the objective seemed to be to replace large amounts of off-line bulk storage, such as magnetic tape. Low storage media costs, on the order of 2.5×10^{-6} cents/bit, are one of the HRMR's strong points.

The HRMR stores digital data in "blocks" by taking the Fourier transform of the bit pattern of that block of data, and forming the interference pattern between this transform and a plane wave reference beam. These blocks, currently holding 484K bits, are stored 62 to a fiche, for a total capacity of 30M bits per fiche. It is therefore possible to hold between one and two magnetic tape reels of data on one or two fiche.

An interesting property of any hologram is that every spot on the hologram's surface contains image information for every spot on the original source. Thus dirt or scratches in the surface of the recording material will not destroy the integrity of recorded digital data to a fatal degree.

In addition to Harris, Honeywell appears to have revived corporate interest in a holographic optical storage device. Details of Honeywell's plans are not yet known.

3.2.5.2 Laser Bit Address Storage Devices

The Precision Instruments Systems Unicon 690 is a currently available write-once, read-only archival memory device that, unlike the HRMR, is directly bit-addressable. Current configurations are capable of storing 1.3×10^{12} bits on-line. The information is "burned" onto rhodium-plated "Data Strips" by a modulated laser beam. The strips are mounted in groups on a rotating drum, which allows an average access time of 13 seconds for a 10^{12} bit memory. Maximal recording rate for requested data is four million bits/second.

The data recording method yields permanent records not subject to read degradation. However, due to the discrete bit recording scheme, the effects of dust and scratches are deleterious to data accuracy. An average error rate of one bit in 10^8 to 10^9 bits is claimed.

The Unicon 690 is supplied with an intelligent controller to minimize the hardware and software impact of interfacing the Unicon to an existing computer system.

3.2.5.3 Electron Beam Addressed Memory

EBAMs show great promise to affect the next "generation" of computers by providing compact, high density storage with storage capacity similar to a large disk, but with access and transfer rates closer to that of semi-conductor memory. Cost per bit of storage is around .02 cents/bit for current EBAM configurations, with potential to fall as low as .001 cent/bit as device density is improved.

The EBAM components are similar to those found in a Cathode Ray Tube. A beam of electrons is generated by an electron "gun" and focused, by means of an electron lens, onto a storage target. The focused beam causes a local charge in the target material corresponding in physical size to the diameter of the focused beam.

"Addressing" is achieved by deflecting the beam to different locations in the target plane. In this manner, either bit-by-bit or block addressing is possible.

To read out from the target plane, the beam is directed at the bit in question and analysis of the velocity of electrons bouncing off the target is made. The presence or absence of a stored charge at the bit location will affect this rebound velocity, causing the bit to be sensed as a "zero" or a "one".

The primary access speed advantage of the EBAM over standard rotational storage devices is the low access time allowed by electrostatic addressing of the target beam. Access time to any position on the target surface is about 80 μsec., as opposed to the multi-millisecond range for rotational storage devices (disks). The readout procedure is, however, partially destructive, so that information must be rewritten after being "read" about 20 times.

First uses of the EBAM technology, which would interface with minimal effort to today's computers and operating systems, will be to replace disk storage. It is possible to build in the necessary refresh logic for the EBAM directly into the controller/interface, so that EBAM operation can be transparent to the host computer. In such an application, the operating system software overhead can effectively "hide" the EBAM's superior access and service time.

Projected, and perhaps more effective use of EBAM is as cache memory, "staging" memory, or as main memory itself. Each of these would have definite impact upon current operating system software, and will perhaps be delayed until such operating system software is developed to take advantage of EBAM's strengths.

Use of the EBAM as cache might be effective if the cache were given a direct path to the CPU, rather than through an I/O channel. Such use would typically replace frequent disk references with cache references, with a corresponding improvement in fetch/read/write times.

Use of EBAM as "staging" memory would involve transferring all information necessary to the execution of a program to the EBAM, whereupon the EBAM would behave as a cache memory with a 100% "hit" ratio. The extension of conventional main memory to include EBAM as well, would allow the inefficiency of addressing EBAM through an I/O channel to be circumvented.

Present Development: Three companies at least are involved in EBAM development: General Electric Research Labs, Stanford Research Institute, and Micro-Bit Corporation. All have working breadboard models, and appear capable of supplying operational EBAM technology in the 1981 timeframe.

<u>Long Range:</u> Access time should improve from the present 80 μsec. to 5-10 μsec. Maximum read rate can be improved from 4 to 8 Mbits/tube. The write rate would be increased to 4 Mbits/tube.

3.2.5.4 **Tunable Dye Laser**

IBM has recently secured a patent on a technique that could be applied to a write-once, read-only storage system. The technique involves precisely controlling the frequency of the laser that projects upon a photoreactive material. It has been shown that the material can "remember" the frequency of the laser that exposed it, and differentiates such exposure at one precise frequency from another such exposure at a second frequency. In theory, it should be possible to expose the same tiny area to a sequence of frequencies and to have the material "remember" each exposure. Thus whole bit patterns could be stored at a single tiny point by assigning each bit in the pattern a unique frequency.

Readout takes place by addressing each point by laser beam deflection, then stepping the reduced-level laser frequency and measuring the absorption of the beam by the photoreactive material. If the material has been "burned" at that frequency, it will not "absorb" the readout beam. Work on this technology is at a very early stage of development, so that the earliest that such technology might be available is the 1985 timeframe. However the storage density potential and rapid addressing of this technology bears close observation.

3.2.5.5 Magneto-Optic Beam Addressed Storage Disks

Honeywell is researching a technique for using laser optics to alter the magnetic properties of a manganese-bismuth film deposited on glass. Using media organized like a conventional magnetic disk, bits are stored

in concentric tracks. Readout occurs by observing the same laser beam at a lower intensity either passing through or being reflected by the film.

The advantage of this technique is extremely high recording density. However, the disadvantages of requiring a powerful laser, long "write" time, and the standard problem of access time being dependent on mechanical "head" movement leave questions about the potential of such technology.

3.2.5.6 Amorphous Semi-Conductors

Amorphous semi-conductors have been the focus of much professional controversy. The basic technology has been known for fifteen years, but has failed, as of yet, to become generally accepted. Recently, However, the Burroughs Corporation signed a non-exclusive licensing agreement with ECD, which may bring the technology closer to the marketplace.

Amorphous semiconductors are a class of glasses (chalcogenide glasses) that contain other elements, most commonly tellurium, arsenic, silicon, and germanium. These materials form an amorphous, glass like structure, that are disordered under ambient conditions. Application of electric energy induces a local ordering in the material, which changes both the materials' opacity and electrical resistance. The area may be "erased" by a short duration, high intensity current pulse.

The primary claimed uses for this technology is in "mostly-read" memories. Strengths of the technology for this purpose are permanence, extremely short access times, high density and direct bit-level addressability. Weaknesses include a definite limit on the number of times an area may be erased, which leads to the classification of "mostly-read" memory.

While the technology has exhibited promise for some time, it has not been brought to the marketplace in a directly usable form. Perhaps the recent agreement with Burroughs will change this.

3.2.5.7 Josephson Devices

Josephson technology, named after Brian Josephson, who first documented the relationship of extremely low temperatures to superconductivity, presents the greatest potential for long-term impact on the nature of computer hardware. The figures for potential computers based on this technology stretch the imagination: switching speeds of ten times today's fastest machines, with miniature logic requiring only one-thousandth the power of present LSI devices. These impressive character-istics come at the cost of keeping the circuitry at $-269^{\circ}C$, which involves immersing all components in liquid helium. This allows removal of all heat sink requirements from components allowing microminiature packaging. Extremely short inter-component connections contribute to the greatly reduced logic switching times.

There is little justification for applying this technology to storage technology alone. In addition to the obvious interface problems, the speed of such a storage device would far outstrip its CPU, and would be largely wasted. Most talk now revolves about a Josephson computer. One theoretical computer design would provide a complete analog of an IBM 370/168, with 16 Mb of main memory, in a package 15 cm^3. Such a device could run 20 times as fast as its 370 counterpart, while requiring only 7 watts of power. However, the cooling device would itself require 15 kw, so the power advantage must be placed in perspective.

Josephson technology will define a new "generation" of computers. Whether it is close enough to define the fourth such generation is a matter of speculation.

3.3   Storage Hierarchy

As has been stated before, no single form of memory has been found
to satisfy all of the storage requirements of computer systems.
Computer developers and users have attempted to compensate for the
difference in speed between internal and peripheral storage. Multi-
programming and virtual storage concepts, among others, have been
invented solely to make up for the speed disparity. A realistic
consideration also is a hierarchy of memory devices.

In order to minimize cost and maximize access throughput, what is needed
is an evaluation of availability requirements and frequency of usage for
each data file, thereby optimizing the data files distribution over the
different storage stages in the system. Figure 3.3-1 illustrates such a
system with storage types ranging from archival to low and high-speed
storage. In general, the more massive storage devices will require longer
access times, but are less expensive. Therefore, as requirements for
accessability increase, data may be transferred to a higher speed,
lower volume, and more expensive storage medium. Cost and speed is
thereby traded for volume. Given a particular application and the
state-of-the-art of available hardware, an assignment optimizing programs
car be developed minimizing cost and response time for an application with
certain task distributions.

Mass Storage Hierarchy

	64K Bytes	2M Bytes	16M Bytes	2G Bytes	200G Bytes
Processor	Processor Cache	Main Storage	Data Base Cache	Disk	Mass Storage
	100ns	800ns	0.5ms	30ms	5000ms
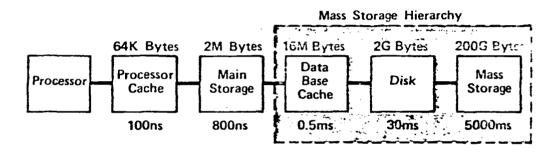
Figure 3.3-1   Storage Hierarchy

By using hierarchical decomponsition, both functional and physical, a
highly parallel information management system can be implemented.
Such a system, called the INFOPLEX (Figure 3.3-2), is presently
under study at the Center for Information Systems Research in the
MIT Sloan School of Management.

In INFOPLEX the storage hierarchy management function is distributed by
means of multiple processors servicing separate physical storage levels.
In this manner, these processors may operate asynchronously and in
parallel. Thus, a request for a lower level function is accomplished
by a inter-processor signal to one of the processors that implement that
level. By incorporating queueing facilities and internal multiprogramming
within each of the processors, a high performance pipeline can be
attained. Although such extensive use of processors has been quite expen-
sive in the past, the advent of low-cost microprocessors makes such a
system economically feasible.
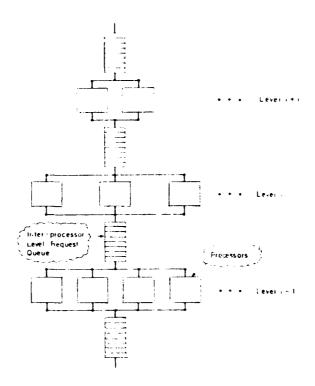
Figure 3.3-2　　　INFOPLEX

3.4 Data Base Management Machinery

New architectural concepts have recently come to the foreground that
promise to bring even more new and fresh life into the data base manage-.
ment environment. Not only have new architectures been proposed in the
academic world, but several have now entered the realization phase in
industry. We will, in this section, first discuss data base accessing
techniques, then several approaches to associative processing data base
management machines, followed by some innovative designs of a rather
different nature.

3.4.1 Data Base Accessing Techniques

Applications which require a great deal of interface with a large data
base must perform extensive I/O. As a result, the throughput of the
host computer system decreases dramatically and the CPU becomes under-
utilized. The data flow through the system can almost always be increased
by reducing data access time.

An in-depth study of anticipated productivity (useful records per
second) requirements might show that random access techniques can still
perform adequately. It might also be discovered that the bulk of record
demand volume exists in easily identifiable requests for large portions
of the data base. In this case, if the data were sequentially stored, all
other requests could be locked out to enable a physically sequential
read-out. This approach, however, is heavily application oriented and
requires a high level of rigidity in the design of the data base.

When contemplating a large, multiuser system, absorbing updates at a
near real-time rate from varied sensor sources, employing functions
requiring extensive access to large portions of a very large data base,
and supporting satellite systems with frequent data base subset down-
loads, it seems obvious that a random access approach to data base
management on disks will be woefully inadequate. For these reasons
other approaches must be considered.

Another, and almost as important, issue to be considered is that of
flexibility. When accessing data randomly it is necessary that control
schemes be utilized to provide succinct access paths to desired data.
Such schemes as index lists, pointers, and hashing techniques employed in
random access processing, require that access paths be identified and em-
bodied in the design of the data base. When using indices, for example,
those fields of a record which will be used often for qualifying queries
must be identified before the data is loaded. A query on an unindexed
field will cause the entire data base to be read. In addition, these indices
must be maintained. It is impractical to index every field, thus causing
a complete duplication of the file and index maintenance overhead.

These techniques are intrinsically inflexible, requiring a complete data
base reorganization every time a requirement occurs for a new index, or
chain of pointers. The relational model of data does not require these
control mechanisms. Basically, all of the control is in the data itself,
thus making possible dynamic creation of relations. If a new relation or
access path is requested, no reorganization or modification of the data
base is required. Similarly, if new fields are added to the data base,
the original data is unaffected.

A-36

It would seem advantageous, if practical, to implement a relational
data base. The mechanics of a relational system require that the data
be read sequentially, from beginning to end. The only way to do this
efficiently on disks is to stream the data, accessing it physically
sequentially, at data transfer rates, thus avoiding seek and latency
penalties.

3.4.1.1 The Data Streaming Architecture

The basic principle of data streaming is that a marked improvement in
data throughput can be achieved by reading disks sequentially rather than
randomly, thus avoiding mechanical arm positioning and rotational delays.
This improvement is on the order of 50 to 1 for standard disk technology
and 1000 to 1 for simultaneous multiple head read disks supported by
parallel and enhanced processors able to keep up with such a prodigous
flow of data. It should be pointed out that such processors do currently
exist and that a nine head simultaneous read disk has recently entered the
marketplace, representing a 450:1 improvement. The advantage is useful,
however, only if the proportion of the data base actually needed is larger
than the inverse of the advantage. That is, if the advantage is 1000 to 1,
then more than 1 segment in 1000 that are delivered must be useful to
someone. This seems a not unrealistic situation considering the many users
expected for intelligence data bases, and the proliferation of intelligence
system functions which may often require the entire data base.

An added benefit of the data streaming approach, based on the relational model of data, is the marked simplicity of access techniques as opposed to the complexity of random access software techniques. A data streaming approach is relatively straightforward, consisting basically of read-examine operations. Current random access techniques are highly complex procedures involving the use and maintenance of such logical structures as index lists, pointers, etc.

The single apparent drawback, though not, it would seem, a serious one, is the underlying limitation on speed of response. The 1000:1 technology can process a 300MByte data base in, roughly, 20 seconds; average response for a simple query for which an answer is found will be 10 seconds. This effect can be mitigated by several means: increasing the number of heads on an arm, decreasing the size of disks while increasing their number, and increasing the number of surfaces. All of these solutions are limited in theoretical potential though they may be perfectly adequate for practical applications in intelligence data processing. Smaller but more disks, for instance, become burdensome after an improvement of, say, 4 to 1. This would mean that a 3 billion byte data base which resided on 10 spindles and provided 10 sec. avg. response, would now reside on 40 spindles and provide 2.5 sec. avg. response, which is very good for most applications but may not be for all. Also, this average response assumes that criteria are internal to the data item and not external, or indirect. For every level of indirect criteria, such as occurs in advanced correlation and dynamic relational systems, a full pass of the data base, 5 seconds, would be required.

While the inherent delay of 5 seconds per pass for the 40 spindle system, and 20 seconds for the 10 spindle system is characterized here as a disadvantage, it is so only relative to what one may wish. These responses are a marked improvement over existing systems which employ random access data management schemes involving pointers and chains and typically require 3 to 4 minutes to answer a three level indirect query on a relatively small data base. It should also be pointed out that such responses were made possible only by extensive pre-production effort in

A-38

the form of extensive data base design. Queries against data in
current systems must be provided for in advance by creating all
necessary links between data items prior to or at data load time. In
a data streaming system these logical links can be dynamically created
as needed. Therefore, response times quoted are for any criteria which
may be requested by the user and which are present as data in the
data base.

Some thought, of course, must be given to the effect of update functions
on this design. Since a block must be read before the decision to
update it can be made, an update will remain in the system for one
complete cycle of the data base after the decision to write. This will
require that memory be provided to hold the updated block while it waits
for its appropriate position to again come under the head. If this memory
is large enough only for one block, only one block can be updated per
cycle. This might be mitigated by separate write heads, trailing the
read heads. Probably the best course would be to provide auxilliary
memory dedicated to holding update block images for writing through a
common read/write head. As many blocks could be updated per cycle as
could be held in this memory. Such a design, coupled with generalized
control software would be flexible enough to allow the easy addition of
more memory if a heavier update load were to be accommodated. These
"solutions" are offered merely to indicate that solutions are at hand
and that no great hurdles are anticipated in solving the update problem.
Any difficulty is much more than offset by the removal of "overhead"
update requirements such as maintenance of indices or pointers.

A commercial product doing string matching on-the-fly is Operating
System's AFP (Associative File Processor). The AFP is characterized
by a parallel search system handling up to five 200 Mbyte disks whose
data is sifted for hits in the Associative Crosspoint Processor against
8192 bytes of parallel keys (Figure 3.4.1.1-1).

Parallel queries are allowed and are kept track of in the Computer
Crosspoint Processor (CXP, a PDP-11 in the single data channel AFP).
This option is especially effective when many users are querying the
same data set. Some overlap is present since the AXP can be searching
for a match independently of the CXP, while that processor is keeping
track of prior matches. The system is specialized toward searching
unformatted files, and seems to be less effective in multi-level
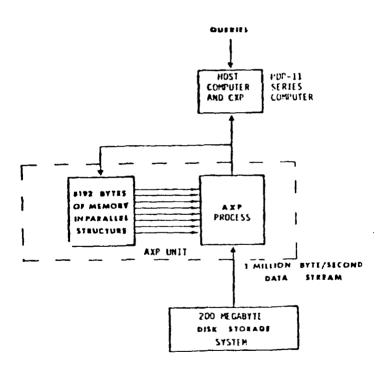searches because of its one-way-street approach. [BIRD77].

Figure 3.4.1.1-1   Single Data Channel AFP

A-40

3.4.1.2 The Hybrid Data Base Machine

The hybrid architecture is an attempt to apply both sequential and random techniques in one system to address both throughput requirements and response requirements of an unusually demanding nature. It is composed of two distinct partitions, random and streaming processing, and an optional adjunct to the control processor which collects and processes statistics concerning data base access activity. The function of the statistics processor, if implemented, would be to aid the central processor in choosing data access paths. (See Figure 3.4.1.2-1).

It is the intent of this organization to route requests which do not require immediate attention, such as updates and data base subset downloading functions and normal queries which require no better than 5-10 second response, to the streaming processor; and fast response queries to the random processor. Further distinction can be made between queries in that certain types may actually be faster if streamed. The selection of the route to be taken can be simple (predefined) or complex (algorithmic). Regardless of the selection procedure, the software for this architecture is more complex than pure data streaming in that it includes both data streaming logic and random access logic, as well as control logic, however simple.

This architecture also requires specially designed disks. Disks must be built which have one read/srite head for sequential processing and one read-only head for random processing connected to a separate controller.
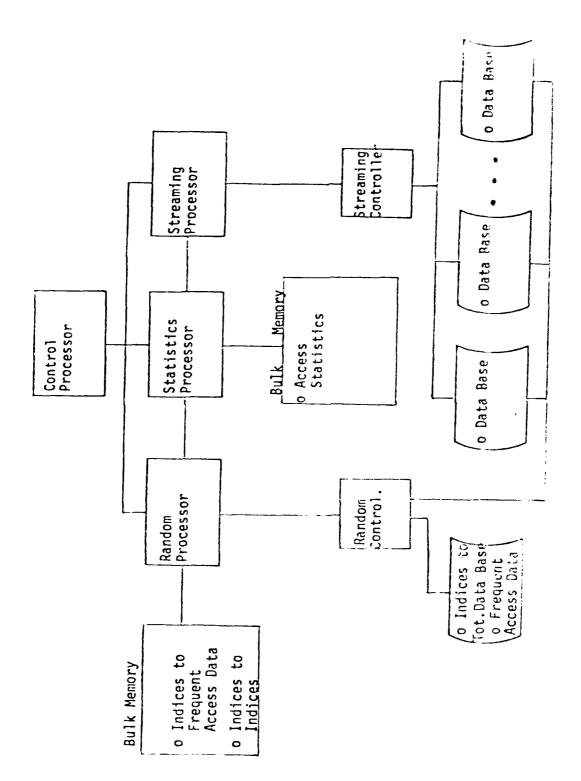
Figure 3.4.1.2-1   Hybrid Data Base Machine

Thought, again, must be given to update activity. It may be performed
by the streaming processor as treated above. Another possibility presents
itself - that of passing updates to the random processor once the
identification of the blocks to be updated has been made and the update
images have been established through streaming locations and control
processor modification. This would provide a possible approximate 20 block
updates per second, but would interfere with the retrieval activities
of the random processor.

It can be seen in the picture of the hybrid machine that bulk memory is
provided as an added resource to the random processor. This bulk memory
can contain the first level or levels of a multi-level, hierarchical index
to the total data base. The more levels contained in the bulk memory, the
fewer disk reads are required to find the pointer to the desired record.
Worthy of much more study are the storage hierarchy schemes, such as disk
cache, applicable to random processing, which keep copies of small but
often accessed portions of the data base in relatively expensive and small,
but very fast storage media, such as CCDs, or bubble memories. A cache
approach maintains this distribution automatically as a result of actual
accesses and requires no decision on the part of the user or data base
administrator. It is in just such ways that the current state of the art
of alternative, advanced memory systems may be applied to enhance total
system power and speed.

3.4.1.3 Concurrent Streaming Architectures

To process data at disk transfer speed requires that a minimum number
of instructions be performed on each item of data. The decision to
"keep" a data item must be made before the next item appears. Since
each item must be examined with respect to possibly many requests, it
is not possible to process at disk transfer speed on a single, sequential
processor. Three approaches to the problem will be described.

A-43
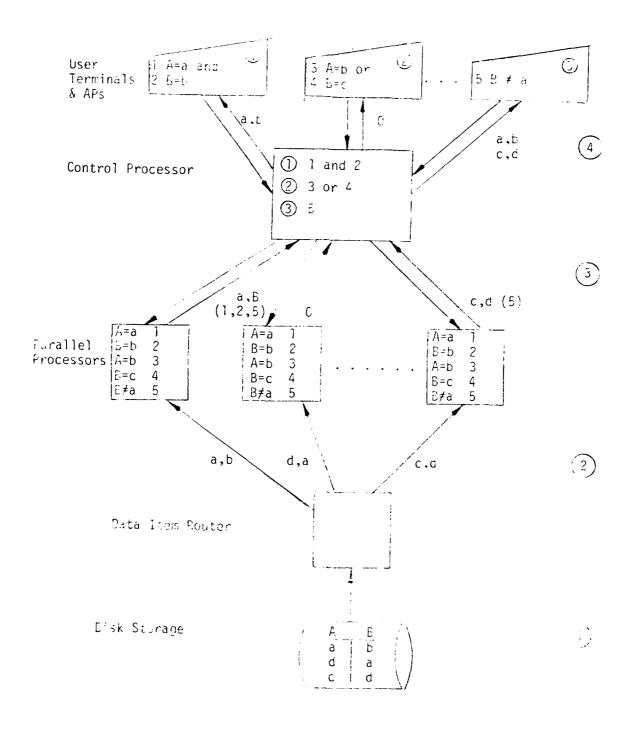
o   Parallel Processing on Multiple Sequential Processors.

In this approach, multiple parallel sequential processors
are filled by the control processor with all current
queries. Data items are parcelled out to the processors
on a round robin basis. A given data item passes through
only one of the parallel processors. The number of
processors required to keep up with the disk is the
execution time for a complete examination of a data item
divided by the time required to read the item from disk.
Complex queries are broken down by the control processor
and passed to the parallel processors as simple queries.
Upon return of hits, the control processor recombines
the queries and performs ANDing and ORing functions.
(See Figure 3.4.1.3-1)

(1)   The data items are read from disk into the Data Item
      Router.

(2)   The DIR sends items to parallel processors for simple
      query selection.

(3)   The parallel processors send items which qualify on
      any of the simple queries to the central processor.
      The identification of each query which qualified the
      item is also returned.

(4)   The central processor performs ANDing and ORing of
      results and sends qualifying items to appropriate
      requestors.

o   Pipelined Processing on Multiple Sequential Processors
     In this approach, multiple processors are each loaded by
     the central processor with portions of queries. Data
     items are passed through all processors sequentially,
     each processor performing a subset of the qualification
     operations. Other characteristics of this configuration
     are similar to those of the previous configuration. (See
     Figure 3.4.1.3-2).

o   Associative Processing

     In this approach, the sequential parallel processors are
     replaced with an associative processor. The control
     processor, just as in the other approaches, breaks down
     complex queries into their simple components, which it
     then moves to the associative processor memory. The
     incoming data stream is passed through the associative
     comparand register and passed against all of the queries
     simultaneously. Hits are returned to the control
     processor which completes the complex operations as in
     the other approaches. (See Figure 3.4.1.3-3)

Figure 3.4.1.3-1 Concurrent Processing on Multiple Sequential Process...
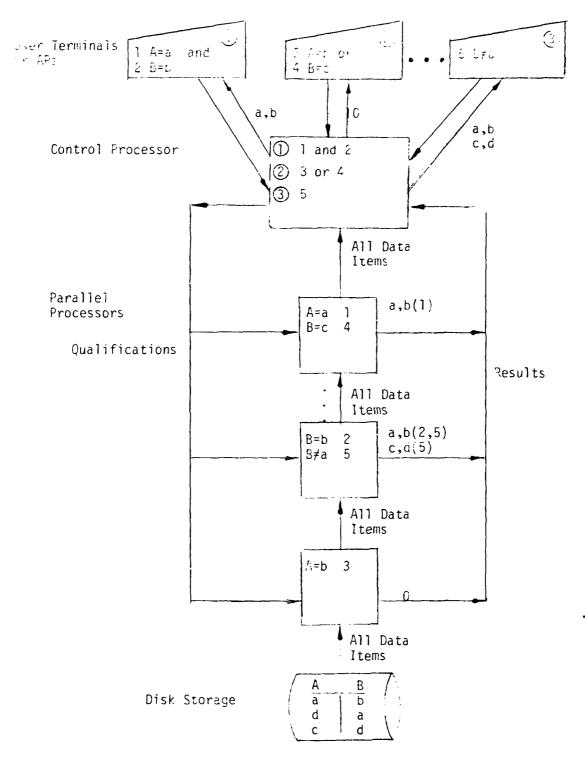
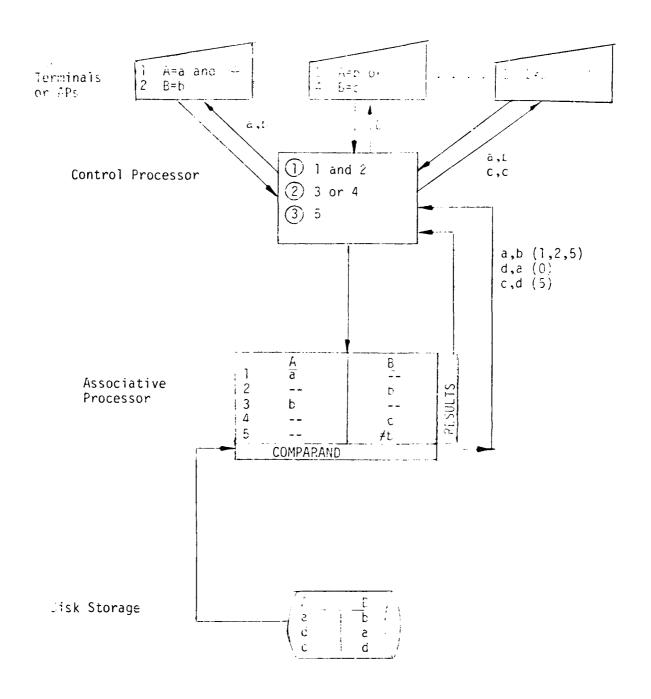Figure 3.4.1.3-2   Pipelined Processing on Multiple Sequential Processors

Figure 3.4.1.3-3 Associative Processing

A-48

3.4.2 Associative Processing

The essence of an associative processor is its ability to locate data items by part or all of their value - content, rather than by location (address). In addition, in current associative processors, arithmetic and logical operations can be performed at the word level in parallel. There is one major advantage to an associative processor - its speed of operation when applied to searching of memory for a specific item. To get comparable speed for catalog look-up in a non-associative processor could, depending upon the types of data involved, could take considerably more ingenuity. Since no indices need be maintained, update is also simplified.

Figure 3.4.2-1 shows in a small example the concept of a parallel match resulting in a result mark in the search results register. Mask register and word select registers are of use in the selection and subsequent compression of the data set to those fields and words that are of interest to the instruction in question.

Parallel processing as of 10 years ago was too expensive. The processors themselves were bulky, power hungry and difficult to manufacture and maintain because of the large part counts and the methods used to interconnect the components. Logic speed restricted the complexity of command/data multiplexing that could be obtained without unacceptable execution times. The provision of multiple processing elements with adequate logic was extremely difficult and the necessary interconnections were virtually impossible to make except for small configurations. The resulting small size of associative memories caused the associative processor to be of limited advantage when there was a large mismatch between the associative store and the total data base. In these cases most time was spent on I/O.

	CONTROL UNIT				MATCH INDICATOR	
0	35	0		DATA REGISTER		MULTIPLE MATCH RESOLVER
0	1	0		MASK REGISTER		
JOHN	30	1	1	0		
JIM	20	2	1	0		
BEV	10	3	1	0		
SALLY	35	8	1	1		
BILL	5	4	1	0		
ART	90	6	1	0		
DOUG	20	7	1	0		
LARRY	10	8	1	0		

GENERAL FORM OF THE FIELD DESCRIPTOR

NAME	DAILY SALARY	EMPLOYEE NUMBER

WORD SELECT REGISTER
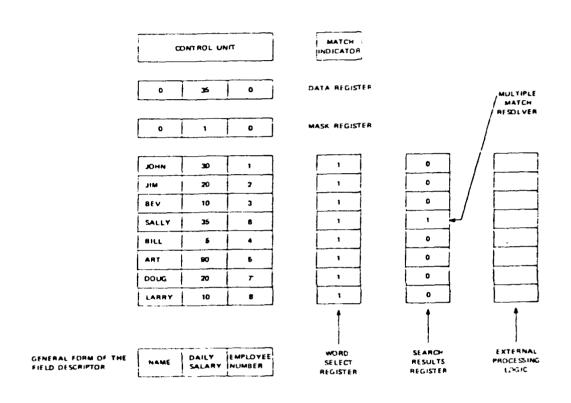
SEARCH RESULTS REGISTER

EXTERNAL PROCESSING LOGIC

Figure 3.4.2-1 Associative Processor

A-50

The situation has noticeably changed in just a few years. Discrete
component logic has rapidly given way to ever increasing complexities in
large scale integrated circuitry. Semiconductor memory has come of age
with its high speed (under 100 nanosecond cycle time) low power require-
ments, high bit density, and integrated addressing logic. Super high
speed bipolar devices introduce 10 nanosecond switching times as off-the-
shelf capability. Thousands of connections are routinely made within a
few square millimeters. It is now possible to package a complete computer
(CPU, memory and timing/bus control) in less physical space than pre-
viously required for the memory alone. In other words, the electronics
industry has finally achieved those capabilities that it needs to implement
the types of parallel and associative processing architectures considered
more than ten years ago.

The SIMD (Single-Instruction-Multiple-Data streams) category, containing
parallel as well as associative processing machinery, can be described
as follows [HIGB73]:

o   Parallel processor - a processor that processes data in
    parallel and addresses data by address instead of by tag
    or value.

o   Associative memory (processor) - a processing device that
    operates on data addressed by tag or value rather than
    by address.

o   Associative processor - a processor with two subsystems -
    an associative memory subsystem and a serial processor
    subsystem - which share a common memory.

o   Parallel associative processor - a processor that is
    associative and also operates on arrays of data.

For data base processing purposes we will be mainly interested in the
form of SIMD processing mentioned last, the parallel associative
processors. The following distinctions can be made in this subcategory:

A-51

o   Fully Parallel Associative Processor:

        -       Distributed Logic Associative Processor (OSI's AXP)
        -       Parallel Element Processing Ensemble (PEPE. etc.)

o   Slice-Serial Associative Processors (STARAN, OMEN, etc.)

o   Disk-Oriented Associative Processors

This mapping into subcategories is not unique and we will find that
particular machines can be categorized in more than one way.

The common factors in these respective implementations is that they
all try to solve the problem of a virtual-to-physical-system-mapping
within the technological constraints, a limiting factor that they are
all subject to. It can even be conjectured that the development from
research into fully parallel, via slice-serial, to disk-oriented
associative processors is a result of the progressive understanding of
the limitations that technology inescapably will introduce in data
base management systems.

In the following sections we will highlight the various forms of parallel
associative processing, especially as they relate to new developments
in computer architecture for data base processing systems. Older
machinery in the three subcategories have been described extensively ir
Yau and Fung's [YAUF77] survey of Associative Processor Architecture.
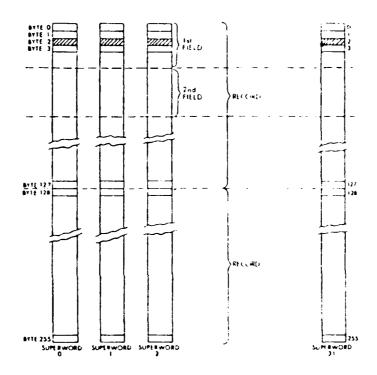
3.4.2.1 Fully Parallel Associative Processors

3.4.2.1.1 Distributed Logic Associative Processor

Of old this form of associative processing has been considered to be
cellular logic systems constructed from cells implementing associative

functions of various complexities. Kautz's cell for the implementation
of an AC4M (Augmented Content-Addressed Memory) system is a prime
example of this approach [KAUT71]. Currently two approaches along those
lines, but at a higher level of complexity, appear realizable from
the technological point-of-view:

o   Semionics' REM (Recognition Memory, [LAMB78]. An Add-in
     Associative Memory Module with a capacity of 4k bytes allowing
     for a modular design of an associative memory. The memory
     module is organized as 16 superwords of 256 bytes. (Figure
     3.4.2.1.1-1)

o   Brunel University's Associative Processor Chip. [UNKN78].
     The Micro-APP single-chip, to be implemented in Plessey's
     proprietary I^2L process, is still in the development stages.
     A test-chip for a 16 byte content - addressable memory, a
     100 mil-square chip in a 40-pin package, will be checkpoint
     along the development (Figure 3.4.2.1.1-2).

Some attempts have been made to produce a cellular set-up of an associative
memory, noticeable among them is the ALAP (Associative Linear Array
Processor) developed at Hughes Aircraft [FINN77]. In it, words, consisting of
short shift registers and contained in a cell with arithmetic and
associative capabilities, are strung together to form a large associative
processor (Figure 3.4.2.1.1-3). The basic module, containing a 112
64-bit words, is implemented on one LSI-chip. The approach is still
that all cells together form an associative memory; currently ALAP
designers are moving towards a system containing, per cell, 64K byte CCDs
rather than 64-bit words, thereby turning it into a segment-sequential
memory such as will be discussed in Section 2.3 Those systems in which
a single data stream passes through a relatively complex cellular system
handling one or more complex queries are still somewhat distributive.
Mukhopadhyay's work on hardware algorithms for nonnumeric computations
[MUKH78] has that flavor. The cellular structures that he developed are
intended to be part of a machine facilitating effective execution of the
SNOBOL language.

A-53

BYTE 0
BYTE 1
BYTE 2
BYTE 3

1st FIELD

2nd FIELD

RECORD

BYTE 127
BYTE 128

RECORD

BYTE 255

SUPERWORD 0   SUPERWORD 1   SUPERWORD 2

0
1
2
3

127
128

253

SUPERWORD 31

Layout shows 8k of REM (two 4k
boards) with two equal length records/
superword, and records divided into
fields. Shaded bytes are accessed in
parallel in REM operation if byte
position 2 is specified by low order
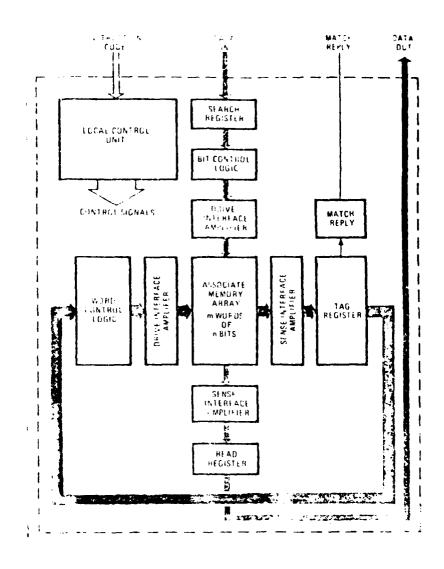address bits.

Figure 3.4.2.1.1-1 Semionics' REM

A-54

Figure 3.4.2.1.1-2   Brunel U's One Chip Prototype of Associative Parallel
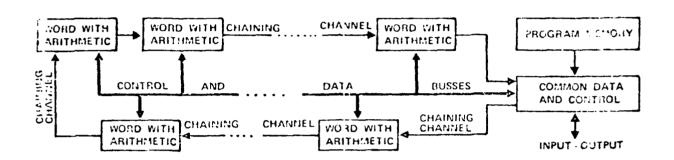   Processor

A-55

Figure 3.4.2.1.1-3   ALAP Configuration

A commercial product doing string matching on-the-fly is Operating
System's AFP (Associative File Processor) [BIRD77].

3.4.2.1.2 Parallel Element Processing Ensemble

PEPE is a parallel-associative processor developed for the U. S. Army
Ballistic Missile Defense Technology Center (BMDATC), in Huntsville, Alabama
[MART77]. Various organizations have been involved with its design and
implementation, among them Bell Telephone Laboratories, System Development
Corporation, Honeywell and Burroughs. PEPE consists of three sequential
processors - arithmetic control unit, associative output control unit,
correlation control unit - controlling typically 288 processing elements
operating in parallel. Each parallel element in the ensemble contains a
correlation unit, an arithmetic unit and an associative output unit, as well
as a 1 k local memory.

Figure 3.4.2.1.1-4 illustrates the PEPE system organization. It is note-
worthy that execution overlap exists in every processing element, so that
a 288-element PEPE can effectively execute up to 864 instructions simul-
taneously. In addition it should be noted that this system provides
general arithmetic and correlational arithmetic hardware as well as
associative functions per element. The total number of elements in the
ensemble is only restricted by cost, making this a machine with growth
potential.

Somewhat simpler parallel elements were applied in HAPPE (Honeywell
Associative Parallel Processing Ensemble, [MARV73] and in RAP (Raytheon
Associative/Array processor, [CONR74]. These systems followed the same
array processing philosophy as PEPE and curiously enough they, as well
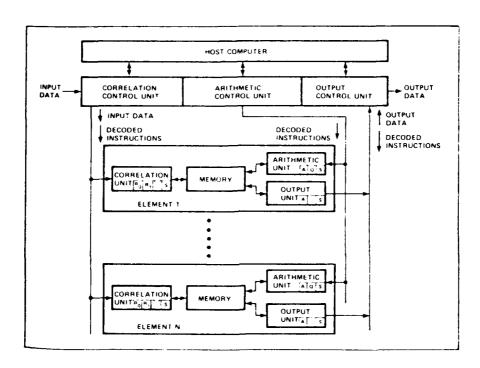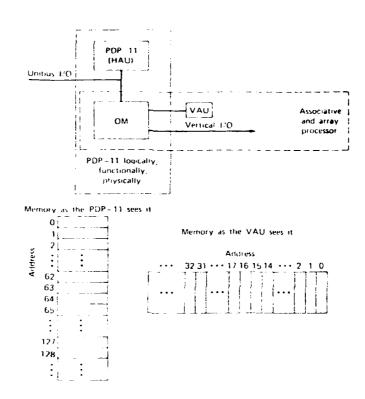as PEPE, originated from radar tracking research.

Figure 3.4.2.1.1-4   PEPE System

3.4.2.2 Slice-Serial Associative Processors

So-called vertical data processing has been introduced in Sanders Associates'
OMEN computer and in Goodyear Aerospace Corporation's STARAN computer. Both
machines center around a PDP-11 central processor. The OMEN computer
[HIGB72] as well as its successor, the SCAT [BURR77], in order to
facilitate vertical data processing, contains a so-called orthogonal memory
(Figure 3.4.2.2-1), which, to the sequential PDP-11 processor, looks like an
ordinary serial memory. At the same time, to the associative and array
processing units, the memory looks transposed. Bit slices over multiple
PDP-11 words then become words to the associative and array processing units
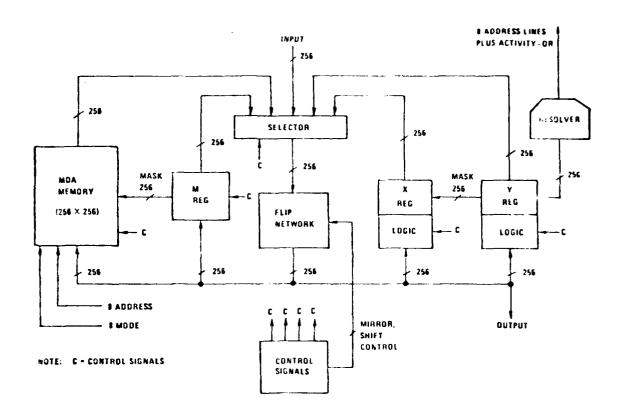and vice versa.

The SCAT - 16 system comes with up to 4, 4 track parallel, 300 M byte moving
head disks, a 17k byte orthogonal memory, and vertical arithmetic unit (VAU)
consisting of a 16 by 8 data key matrix of processing elements and additional
control elements to resolve the findings of the 128 element array.

The STARAN [RUDO72] consists of a PDP-11 as central processor, as mentioned,
and up to 32 separate associative processing modules. Each module (Figure
3.4.2.2-2) contains a 256 word by 256 bit mixed-mode access memory, an
alignment network, and several bit-slice registers implementing the associative,
arithmetic and logical functions. The mixed-mode access memory results in
an extensive addressing capability, allowing 256 different ways of accessing
the array element's memory. Vertical data processing will perform an
associative function by operating on bit-slices over several words. A
recently announced model E is quite similar to the original STARAN. In

Figure 3.4.2.2-1   OMEN-60 Memory

Figure 3.4.2.2-2   STARAN Array Module

addition to some changes in technology, the associative processing module contains 256 words of 9216 bits, although processing is still over 256 words by 256 bits.

Wolfgang Handler [HAND74] has proposed to take these concepts one step further. After observing similarities between horizontal and vertical data processing hardware he suggested to allow common memory to contain information of both types and to adjust the processing accordingly. These notions and the unification they bring to data processing are the center of much attention at the University of Erlangen, Nürnberg, West Germany. One of the objectives of this work is to be able to define a machine containing both types of data using conventional hardware.

3.4.2.3 Disk Oriented Associative Processors

In this section we will describe systems in which the mismatch between virtual and physical associative processing capabilities is resolved by performing the associative function at the I/O interface between secondary and primary storage. In most cases this takes the form of concurrent operations on multiple disk tracks. So rather than following the staging philosophy of Figure 3.4.2.3-1 with its limited I/O performance one would like to increase the total data flow by introducing concurrency of disk operations (Figure 3.4.2.3-2).

Proposals implementing the associative function through disk-oriented devices have been made all along. A typical example is the AFP (Associative File Processor) developed by General Precision, Inc.[CASS67], illustrated in Figure 3.4.2.3-2, in which multiple heads load a small associative memory in parallel, subsequently followed by handling of matches in the general purpose host computer.

More recently there has been a revitalization of the concept of logic per-track devices. Forerunners in this approach have been two systems: CASSM developed at the University of Florida, and RAP, developed at the University of Toronto, Canada.

CASSM [COPE73], Context-Addressed-Segment-Sequential Memory, presumes the data base to be segmented over many tracks and/or disks, thereby resulting in short parallel search times (Figure 3.4.2.3-4). Every head of every track is augmented with a (micro-) processor for auxillary pre-processing. The total track will be treated by the logic-per-track processor each disk revolution (Figure 3.4.2.3-5).

Figure 3.4.2.3-1 Associative Memory as an Accessory Data
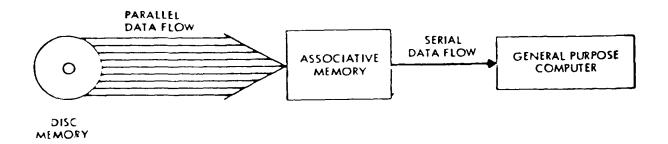Search Device

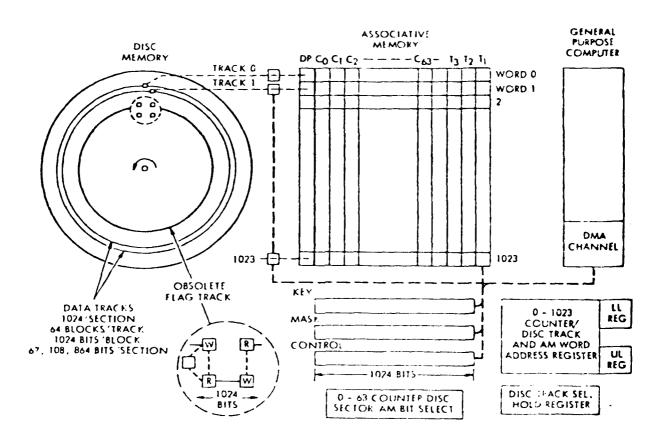Figure 3.4.2.3-2   Associative File Processor Structure
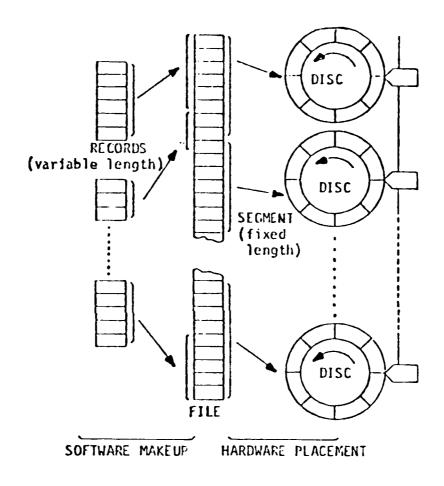
Figure 3.4.2.3-3   AFP Major Hardware Elements

Figure 3.4.2.3-4   Storage of Records as Segments
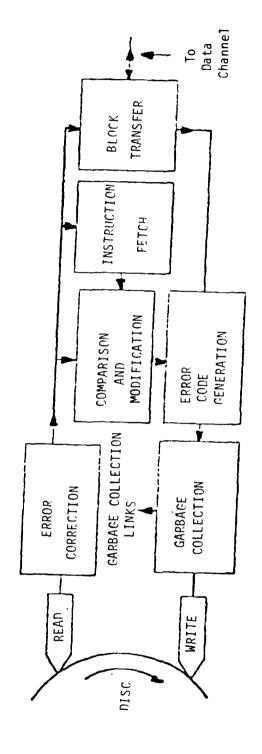
Figure 3.4.2.3-5   Segment Cell Hardware

A-68

A marking mechanism for nested searches, etc. is implemented by a
one-bit wide RAM addressed by the location counter of the tract
segment. Although the design is intended for fixed-head floppy
disks, it can also be implemented with CCD or magnetic bubble memories.
The CASSM system lends itself to various forms of data base structures
including relational systems.

RAP [SCHU78], though conceptually similar to CASSM, is intended
exclusively for relational data base designs, as its name, Relational
Associative Processor, indicates. The system consists of a controller,
a statistical arithmetic unit and a chain of cells (Figure 3.4.2.3-6).
The controller performs the translation and scheduling of high-level
language query directives into the appropriate cell actions. The
statistical unit collects statistics such as average, minimum/maximum
value, etc. The cell itself contains a rotational memory and a processor
to implement the data base functions. The current implementation of
RAP calls for cells each containing 1 megabyte of CCD shift registers,
and a total of between 10 and 100 cells per system.
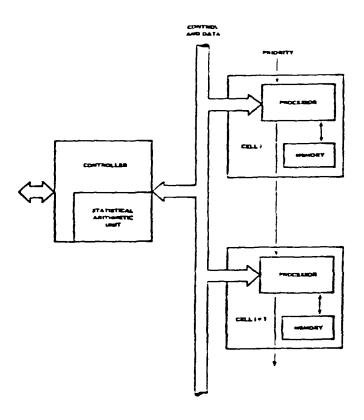
Figure 3.4.2.3-6 RAP.2 System Architecture

After observing the output techniques in STARAN and CASSM/RAP, researchers at the University of Utah decided to combine the principles of CASSM/RAP with an "orthogonal" layout of data on the disk. The system, thus created, called RARES (Rotating Associative Relational Store [LINS75] contains search-cells. They are, however, handling a set of tracks, a band (Figure 3.4.2.3-2). The layout over the tracks allows a high output rate of selected data items provided that there are sufficiently many tracks per band. Assignment of search-module to track-set can be done under program control.

With the coming of age of CCD shift registers and bubble logic memories, many systems are being proposed which employ them instead of disks. An outgrowth of the ASP system, developed by Hughes Aircraft [LOVE73], follows this philosophy in which many short shift-registers perform the associative function when supplied from large shift register bulk storage through a switching matrix. It is also possible to hook the associative shift registers up into longer shift registers, thereby adapting the memory to the application.
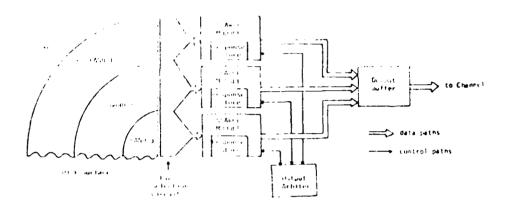
Figure   3.4.2.3-7   RARES Content-Addressing Mechanism

A product also developed along the concept of segment-sequential associative processing is the ECAM [ANDE76] (Extended Context Addressed Memory) produced by Honeywell. In its most recent form, ECAM consists of up to 250,000 words of 4K bit CCDs and appears to the host as a standard high-speed peripheral (such as a disk) (Figure 3.4.2.3-8).

Each word has associative and arithmetic functions, though only localized to that word, so that operations on the total file of 250,000 by 4K bits are more complicated (Figure 3.4.2.3-9).

Still under development, but already patented, is a bubble logic chip proposed by IBM researchers that in addition to the bubble storage will contain all the necessary logic to do searches on that particular segment of the relational data base [CHAN78].
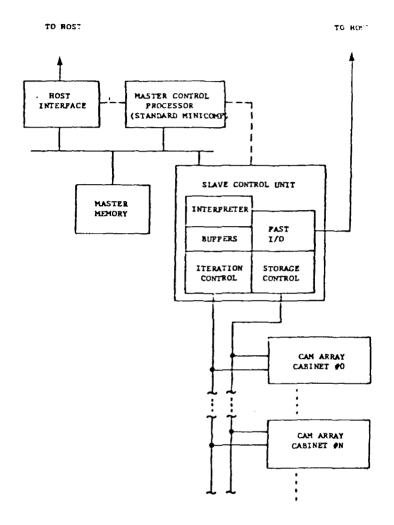
Figure 3.4.2.3-8   Extended Content Addressed Memory
(ECAM) System Block Diagram
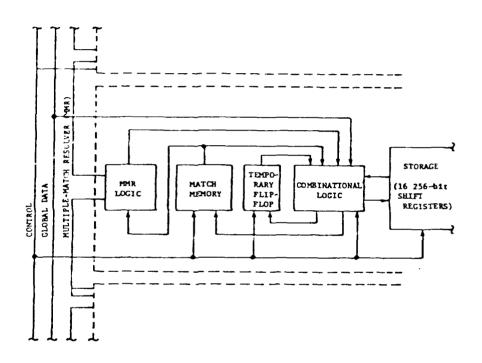
Figure 3.4.2.3-9   Word Structure in CAM Array of the ECAM

3.4.3 Other Data Base Management Machines

The DBC (Data Base Computer) [HSIA77] developed at Ohio State University
is unique in its approach to data base processing. It is consistent
in its objective to implement in hardware those aspects of data base
processing that of late have been part of special data base processing
software. It consists of two loops of memories and processors, the
structure loop and the data loop (Figure 3.4.3-1). The structured
loop has four components:
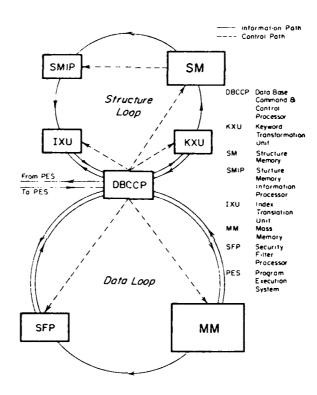
o   Keyword transformation unit (KXU)

o   Structure memory (SM)

o   Structure memory information processor (SMIP)

o   Index translation unit (IXU)

These four elements can operate concurrently, thereby realizing a pipeline.

Keywords are sent to the KXU, whose output is sent to the SM which
retrieves index terms for the transformed keyword predicates. These
index terms are sent to the SMIP, whose output is interpreted by the
.IXU. Output of the IXU to the DBCCP then closes the loop. The data loop
has three components:

o   Data Base Command and Control Processor (DBCCP)

o   Mass Memory (MM)

o   Security filter processor

Queries originally are received by the DBCCP, and subsequently decoded
in the structure loop, results of which are sent to the MM. The outputs
from MM are finally filtered according to security requirements.

Figure 3.4.3-1 The Ohio State Data Base Computer

Both SM and MM are implemented using partitioned content addressable memory technology, somewhat similar to those applied in the RAP, etc. The design calls for certain modifications to existing moving-head disk systems such as the association of a processing element with each read/write head and the concurrent activation of all read/write heads available in the access mechanism.

It is expected that the SM will be considerably more prone to updates than the MM. SM is thought to be of the order of $10^7 - 10^9$ bytes, while MM could be between $10^9 - 10^{10}$ bytes.

The separation of data base control information from the data base itself seems to be a worthwhile decision. The total DBC concept has received wide attention and UNIVAC has initiated an effort to build a machine of this kind.

The machine discussed above clearly is of the MIMD (Multiple-Instruction-Multiple Data Streams) type since concurrent pipelines exist even between the two loops.

A MIMD in the truest sense of the word has been suggested at the University of Wisconsin. The system, DIRECT [DEWI78], consists of a set of LSI-11/03 microprocessors accessing through a cross-point switch a set of CCD memories (Figure 3.4.3-2). In this system it is possible for several queries to be handled at the same time, or even to have several processors handle a more complex query. On a larger scale a similar structure (Figure 3.4.3-3) was proposed by Gaertner [GAER76], the G-471 centered around a cross-bar switch, in which several enhanced PDP-11/45s are searching, in parallel, banks of CWS (Central Working Storage). Though originally conceived as a signal processing machine, it was re-worked into a data base processing machine with the added capability of being a number crunching machine.
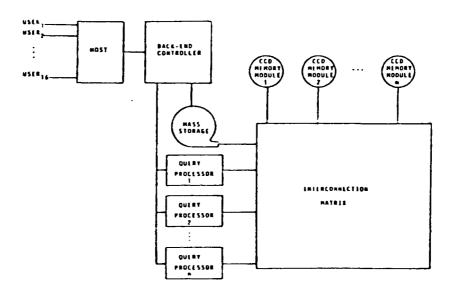
Figure   3.4.3-2   Direct System Architecture

CONTROL-COMPUTER COMPLEX

SEQUENTIAL AND
CONTROL COMPUTER

CRT CONSOLE
PRINTER
PLOTTER

DISKS

TAPES

PROCESSING-ELEMENT ARRAY

PROCESSING ELEMENT

LOCAL STORAGE

PROCESSING ELEMENT

LOCAL STORAGE

REAL-
TIME
I/O
LINES

UP TO 120 KBYTES

4 MIPS

PROCESSING ELEMENT

LOCAL STORAGE

DATA-ROUTING ARRAY

MASS-STORAGE SYSTEM

FAULT TOLERANT
SEGMENTED
PARALLEL ACCESS

MASS-STORAGE
CONTROLLER

DISK
TAPE

MASS-STORAGE
CONTROLLER

DISK
TAPE

MASS-STORAGE
CONTROLLER

DISK
TAPE

MASS-STORAGE
CONTROLLER

DISK
TAPE

TRANSFER RATE
16 MBYTES/SEC/CHANNEL

REAL-TIME
I/O
CHANNELS

CENTRAL WORKING STORAGE

MEMORY
BANK 1

MEMORY
BANK 2

64-BIT PARALLEL ACCESS PER BANK
16 MBYTES DIRECTLY ADDRESSABLE

MEMORY
BANK N

Figure 3.4.3-3   Block Diagram of the G-471 Parallel/
Associative Computer System

A-80

3.4.4 Recapitulation

The field of design of data base machines is very much in a state
of flux. Many manufacturers are working on some kind of data base
system, ranging from the office-environment to very-large-data-
base systems.

Figure 3.4.4-1 illustrates the developments that are taking place
that will eventually lead to the separation of the data base
function from the host. This partition will be advantageous from
a functional and a throughput point-of-view.

It has been shown in this section that a revitalization of associative
processors as they apply to data base processing machinery has taken
place as well as the introduction of new data base accessing techniques.
In particular the emerging of microprocessors, CCD shift registers,
bubble memories and the advancement of disk technology has made this
development possible and worthwhile. We foresee that in the relatively
near future a large range of data base machinery will be available,
potentially to off-load many a burdened data processing center.
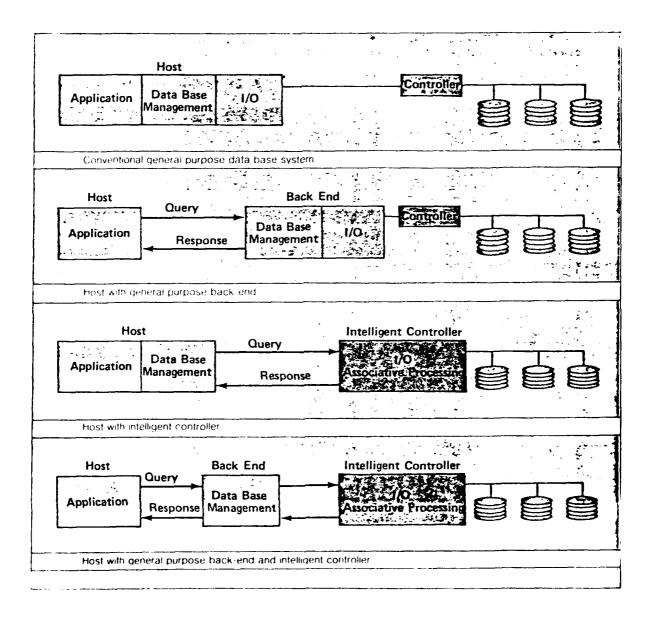
Figure 3.4.4-1   Developments Toward Functional Separation

3.5 Summary

The technology review in this chapter has been concerned with those aspects of current technology development that will make the greatest impact on data base management machinery. Three developments can be discerned:

o   The introduction of new storage devices and their trade-off aspects relative to conventional storage media such as disks, etc.

o   The use of those new devices as well as conventional storage devices to improve current data base management machinery by introducing them as an intermediate medium between CPU and secondary memory, that is as caches, paging devices, etc.

c   The introduction of new data base management machinery concepts resulting from the improved understanding of the storage devices, new as well as conventional.

The trend that is visible through all these developments is the move toward self-contained data base management processors.

3.6 References

[ANAC79]   Anacker, W., "Computing at 4 Degrees Kelvin",
            IEEE Spectrum, May 1979.

[ANDE76]   Anderson, G. A,; Kain, R. Y., "A Content-Addressed
            Memory Designed for Data Base Applications", 1976
            International Conference on Parallel Processing.

[ARNO79]   Arnold, W. F., Durniak, A., "Denser Disk Drives
            poised for Surge", Electronics, June 21, 1979.

[BART78]   Bartolini, R. A., et al, "Optical Disk Systems
            emerge", IEEE Spectrum, August 1978.

[BIRD77]   Bird, R. M.; Tu, J. C., Worthy, R. M., "Associative/
            Parallel Processors for Searching very Large Textual
            Data Bases", Third Wrokshop on Computer Architecture
            for Non-Numeric Processing, Syracuse, New York,
            May 1977, 8-16.

[BULT79]   Bulthuis, K., et al, "Ten Billion Bits on a Disk",
            IEEE Spectrum, August 1979.

[BURR77]   Burrows, J. L., Reynolds, W. E., "Sensor Data
            Correlation System Development", RADC-TR-76-243, July 1977.
            Vol I, AD# A045 541; Vol II, AD#B020 517L
[CASS 67]   Cass, J. L., "Organization and Application of
            Associative File Processors", ONR/RADC Seminar on
            Associative Processing, May 1967, Washington, D. C.

[CHAM78]   Champine, G. A., "Four Approaches to a Data Base
            Computer", Datamation, December 1978.

[CHAN78]   Chang, H., "Bubbles for Relational Database", Fourth
            Workshop on Computer Architecture for Non-Numeric
            Processing, 1978.

[CONR74]   Conranz, G. R., Gerhardt, M. S., Young, C. J.,
            "Programmable Radar Signal Processing Using the RAP",
            1974 Sagamore Computer Conference on Parallel
            Processing, Lecture Notes in Computer Science, 24,
            Springer Verlag, 37-52.

[COPE73]   Copeland, G., Lipovski, G. J., Su, S. Y. W., "The
            Architecture of CSSSM: A Cellular System for Non-
            Numeric Processing", First Annual Symposium on Computer
            Architecture, 1973.

[DEWI78]   DeWitt, D. J., "DIRECT - A Multiprocessor Organization
            for Supporting Relational Data Base Management Systems",
            Fifth Annual Symposium a Computer Architecture, 1978.

[FINN77]   Finnila, C. A., Love, H. H. Jr., "The Associative
            Linear Array Processor", IEEE Transactions on Computers,
            Vol. C-26, No. 2, February 1977, 112-125.

[GABE79]   Gabel, D., "Disc Memories", Electronic Engineering
            Times, August 6, 1979.

[GAER76]   Gaertner et al, "High-Resolution Image Processing on
            a Parallel Computer System", 1976 International
            Conference on Parallel Processing.

[GEOR79]   George, P. K., Reyling, G., Jr., "Bubble Memories
            come to the Boil", Electronics, August 2, 1979.

[GILL75]   Gillis, A. K., Hoffman, G. E., Nelson, R. H.,
            "Holographic Memories: Fantasy or Reality?",
            NCC 1975, AFIPS Conference Proceeding Vol. 44.

[HAND74]   Händler, W., "Unconventional Computation by Conventional
            Equipment", NATO-Defence Research Group Seminar, Design
            and Evaluation of Information Systems, Athens, April 1974.

[HIGB72]   Higbie, L. C., "The OMEN Computers: Associative Array
            Processors", COMPCON 1972, 287-290.

[HIGB72]   Higbie, L. C., "Supercomputer Architecture", Computer,
            December 1973, 48-58.

[HSIA77]   Hsiao, D. K., Kannan, K., "The Architecture of a
            Database Computer - A Summary", Third Workshop on
            Computer Architecture for Non-Numeric Processing, 1977.

[KAUT71]   Kautz, W. H., "An Augmented Content-Addressed Memory
            Array for Implementation with Large-Scale Integration",
            J.ACM, January 1971, 19-33.

[KENN79]   Kenney, G. C., et al, "An Optical Disk replaces 25 Mag
            Tapes", IECE Spectrum, February 1979.

[LAMB78]   Lamb, S., "An Add-In Recognition Memory for S-100 Bus
            Microcomputers - Part 1: Introduction", Computer Design,
            August 1978, 140-142.

Idem, "Part 2: Structure and Specifications",
Computer Design, September 1978, 162-168.

[LAPI73]   Lapidus, G., "Amorphous Devices: at the Threshold",
            IEEE Spectrum, January 1973.

[LEMM77]   Lemmond, C. Q., et al, "Advanced Archival Memory",
            General Electric R&D, Schenectady, NY.

[LINS75]   Lin, C. S., Smith, D. C. P., "The Design of a Rotating
            Associative Array Memory for a Relational Data Bases,
            September 1975, pp. 354-455.

[LOVE73]   Love, H. H., "An Efficient Associative Processor Using
            Bulk Storage", Proceedings of 1973 Sagamore Computer
            Conference on Parallel Processing Processing, pp 103-112.

[MADN75]   Madnick, S. E., "INFOPLEX - Hierarchival Decomposition
            of a Large Information Management System using a
            Microprocessor Complex", NCC 1975, AFIPS Conference
            Proceedings Vol. 44.

[MARI75]   Marill, T., Stern, D., "The Datacomputer - A network
            Data Utility", NCC 1975, AFIPS Conference Proceedings
            Vol. 44.

[MART77]   Martin, H. G., "A Discourse on a New Super Computer, PEPE",
            High Speed Computer and Algorithm Organization, Academic
            Press, 1977, 101-112.

[MARV73]   Marvel, O. E., "HAPPE, Honeywell Associative Parallel
            Processing Ensemble", First Annual Symposium on Computer
            Architecture, 1973, 261-267.

[MEAS78a]   Measurement Concept Corporation, "Architectural
            Alternatives in Data Base Management Machinery", Technical
            Report, October 1978.

[MEAS78b]   Measurement Concept Corporation, "Technology Survey
            (State-of-the-Art)", Working Paper, November 1978.

[MUKH78]   Mukhopadhyay, A., "Hardware Algorithms for Non-Numeric
            Computation", Fifth Annual Symposium on Computer
            Architecture, 1978, 8-16.

[RUDO72]   Rudolph, J. A., "A Production Implementation of an
            Associative Array Processor-STARAN", AFIPS Conference
            Proceedings, Vol. 41, 1972, pp. 229-241.

[SCHU78]   Schuster, S. A., Nguyen, H. B., Ozkarahan, E. A.,
           Smith, K. C., "RAP.2: An Associative Processor for
        s Data Bases", Fifth Annual Symposium on Computer
           Architecture, April 1978.

[SHAP77]   Shapiro, R. M., Millstein, R. R., "NSW Reliability Plan".
           Massachusetts Computer Associates, CA-7701-1411,
           June 10, 1977.

[SPEL75]   Speliotis, D. E., "Bridging the Memory Access Gap",
           NCC 1975, AFIPS Conference Proceedings Vol. 44.

[UNKN78]   Author Unknown, "Associative Processor on a Chip Due
           from British Laboratory", Electronics, January 19,1978,
           o3-o4.

[WARN73]   Warnar, R. B. J., Calomeris, P. J., Recicar, S. A.,
           "Computer Peripheral Memory System Forecast", ICST-
           NBS Special Publication 500-45, April 1979.

[YAUF77]   Yau, S. S., and Fung, H. S., "Associative Processor
           Architecture - A Survey", ACM Computing Surveys,
           Vol. 9, No. 1, March 1977, pp. 3-28.

MISSION
of
Rome Air Development Center

RADC plans and executes research, development, test and selected acquisition programs in support of Command, Control Communications and Intelligence (C³I) activities. Technical and engineering support within areas of technical competence is provided to ESD Program Offices (POs) and other ESD elements. The principal technical mission areas are communications, electromagnetic guidance and control, surveillance of ground and aerospace objects, intelligence data collection and handling, information system technology, ionospheric propagation, solid state sciences, microwave physics and electronic reliability, maintainability and compatibility.

DAT FILM